DOCUMENT RESUME

ED 368 776 TM 021 205

AUTHOR Baxter, Gail P.; And Others
TITLE Analysis of Cognitive Demand in Selected Alternative Science Assessments. Analysis of Structures and Processes Assessed in Science. Project 2.1 Alternative Approaches to Assessment in Mathematics and Science. Cognitive Theory as the Basis for Design of Innovative Assessment.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Nov 93
CONTRACT R117G10027
NOTE 78p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Cognitive Processes; Cognitive Psychology; Cognitive Structures; Concept Formation; *Educational Assessment; Educational Innovation; Elementary Secondary Education; Identification; Knowledge Level; Pilot Projects; Problem Solving; Protocol Analysis; Science Education; *Science Tests; *Scoring; Test Construction; Test Interpretation; Test Use; Thinking Skills
IDENTIFIERS *Alternative Assessment; *Cognitive Demands; Concept Maps; Performance Based Evaluation

ABSTRACT

Analyses of the cognitive activity displayed by students on existing innovative assessments in science were conducted. The project has worked with pilot alternative assessment programs in Connecticut and California to document the match or mismatch between the skills and processes that the assessment is designed to tap and those actually elicited. Exploratory investigation, conceptual integration, and component identification tasks were selected, each varying with respect to grade level, prior knowledge, stage of development, and purpose. Detailed verbal protocols of students in each of the assessment situations in conjunction with observations of student performance, evaluation of student test booklets, and an examination of the scoring criteria provide an empirical basis for linking performance scores with the level and kind of reasoning and understanding. Analysis of these sources of data was guided by a framework of proficient performance that grew from the cognitive psychology literature on the nature of expertise. Focus was placed on the extent to which tasks allowed students the opportunity to engage in higher-order thinking skills (i.e., plan, reason, explain, infer, monitor, or strategically problem solve) and the extent to which scoring systems reflected differential performance of students with respect to the nature of the cognitive activity in which they engaged. Five tables and 19 figures are included. (Contains 28 references.) (Author/SLD)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 368 776

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – November 1993

Project 2.1 Alternative Approaches to Assessment
in Mathematics and Science

Cognitive Theory as the Basis for Design of Innovative Assessment

Analysis of Structures and Processes Assessed in Science

Analysis of Cognitive Demand in Selected Alternative Science Assessments

Robert Glaser, Project Director
CRESST/LRDC, University of Pittsburgh

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

National Center for Research on
Evaluation, Standards, and Student Testing
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

TM021205

BEST COPY AVAILABLE

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

ANALYSIS OF COGNITIVE DEMAND IN SELECTED ALTERNATIVE SCIENCE ASSESSMENTS[1]

Gail P. Baxter, University of Michigan

Robert Glaser and Kalyani Raghavan
CRESST/Learning Research and Development Center,
University of Pittsburgh

ABSTRACT

The purpose of this study is to carry out analyses of the cognitive activity that is displayed by students on existing innovative assessments in science. The project has worked with pilot alternative assessment programs in Connecticut and California to document the match or mismatch between the skills and processes that the assessment is designed to tap and those actually elicited. Three types of assessment tasks were selected—Exploratory Investigation, Conceptual Integration, and Component Identification—each varying with respect to grade level, prior knowledge, stage of development, and purpose. Detailed verbal protocols of students in each of the assessment situations in conjunction with observations of student performance, evaluation of student test booklets, and an examination of the scoring criteria provide an empirical basis for linking performance scores with level and kind of reasoning and understanding. Analysis of these sources of data was guided by a framework of proficient performance which grew out of the cognitive psychology literature on the nature of expertise. We focused on the extent to which: (a) tasks allowed students the opportunity to engage in higher order thinking skills (i.e., plan, reason, explain, infer, monitor, or strategically problem solve) and (b) scoring systems reflected differential performance of students with respect to the nature of cognitive activity in which they engaged.

---

[1] The authors wish to gratefully acknowledge the support and assistance of Joan Baron and Michael Lomask from the Connecticut State Department of Education. They encouraged the trial and pilot study from which we could learn from each other's experience and analyses. We are also particularly grateful to Jennifer Yure of the Pasadena Unified School District for her help in the conduct of this study.

Thanks are extended to Sophie Kesidou, formerly of LRDC, and Mary Sartoris of LRDC for their assistance with data collection. We also wish to thank Tim Breen, University of Michigan, for his assistance with the analysis and his helpful comments on earlier drafts of this report. The opinions expressed, however, are those of the authors and not necessarily those of our colleagues or our colleagues' sponsoring organizations.

Results of this analysis of a diverse range of current assessment practices and scoring systems serve to highlight characteristics of the situations in which students' performances were elicited, scored and cognitive skills assessed. These characteristics or features of assessment provide the basis for the development of a framework for guidance in the development of assessment tasks and scoring systems that link performance scores to levels of student understanding.

Our analyses to date suggest the following results of tasks and scoring. In general, tasks should: (a) be procedurally open-ended affording students an opportunity to display their understanding. Differential approaches to solving the problem are often masked when procedural instructions are provided; (b) draw on subject matter knowledge as opposed to knowledge of generally familiar facts. Indepth subject matter understanding is necessary for generating complex arguments to support an explanation; (c) be cognitively rich enough to require thinking. Recalling specific facts or generating a list of attributes can be accomplished with superficial understanding of a given topic. In contrast, reasoning and problem solving are dependent on knowing when, and under what conditions, specific content knowledge is useful.

Scoring systems should: (a) link score criteria to task expectations. Cognitively rich tasks that ask students to reason with subject matter knowledge demand scoring systems that detail the quality of that reasoning process in a given context. Scoring systems that emphasize easily quantifiable aspects of performance in the end sabotage efforts to measure higher order thinking skills; (b) be sensitive to the meaningful use of knowledge. If students' performances are not to be over- or underestimated, a link between performance score and level of reasoning and understanding is critical. Focusing on key words when students are asked to explain fails to distinguish those students with cohesive knowledge from those with fragmented knowledge, a key distinguishing feature of proficient performance; (c) capture the process students engage in. Knowing when and under what conditions knowledge is useful is reflected in problem-solving strategies and processes. Performance scores that attend to the product without explicit consideration of the process encourage unwarranted assumptions about students' understanding.

These characteristics or features of assessments provide a basis for an assessment development framework that acknowledges proficiency as the ability to use subject matter knowledge to reason and solve problems. Sensitizing test developers to ways in which students' reasoning can be elicited and scored is the first step toward development of assessments commensurate with educational goals.

INTRODUCTION

The development of understanding, reasoning, and problem solving in various subject matters has become a major focus of educational reform. In science, for example, problem identification and representation, the generation of hypotheses, planning and information organization, the exploration of multiple problem-solving strategies, and the evaluation of progress toward problem solution are viewed as critical aspects of scientific understanding. Efforts to develop measures of student achievement commensurate with these educational goals focus on creating assessments that display how students use their knowledge to reason and solve problems in meaningful contexts. The assumption is that assessment tasks can be created, administered, and scored to obtain reliable and valid information about students' higher order thinking skills.

Some attention has been given to the reliability and validity of these alternative assessments (e.g., Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993). Much less attention has focused on whether the tests actually measure higher order thinking skills (but see Magone, Cai, Silver, & Wang, 1992). What is needed is an examination of the kinds of knowledge and cognitive processes that are actually being tapped by these assessments (Linn, Baker, & Dunbar, 1991). Documentation of the match or mismatch between the skills and processes that the assessment is designed to tap and those actually elicited provides empirical evidence bearing on the cognitive demands of alternative assessments.

An understanding of the cognitive structures and activities that underlie successful task completion has been the focus of a large body of cognitive psychology research. Using protocol analysis techniques (cf. Ericsson & Simon, 1984), researchers have characterized the differential performance of individuals varying in subject matter knowledge and skills (competence) in terms of their problem-solving strategies, quality of explanations, and reasoning. Drawing on this research, the project described here has collected detailed verbal protocols of students' performances in pilot assessment situations in Connecticut and California. Analysis of these protocols, in conjunction with observations of student performance and an examination of student answer booklets and scoring criteria, provides a basis for linking performance scores with level and kind of reasoning and understanding.

It would be remiss to analyze the assessment task without the corresponding scoring system that serves to evaluate performance on that task. Clearly if the test is intended to tap certain skills/processes, then students' scores should reflect their level of proficiency with respect to those skills and/or processes. Indeed, one could conceive of a task that taps higher order thinking skills but a scoring system that does not reflect differential levels of proficiency except in the broadest sense (i.e., right or wrong). Innovative assessments, by their very nature, demand innovative scoring systems if the scores are to reflect differential proficiency of students defined by the nature and extent of the cognitive activity they engage in.

Results of the cognitive activity analyses and the study of the assessment tasks and scoring procedures will provide guidance about the ways in which students' performances can be elicited and scored to ensure that appropriate cognitive skills are actually involved (Glaser, Lesgold, & Lajoie, 1987; Mislevy, 1989; Snow & Lohman, 1989). Guidance for assessment development, then, will come from the description of cognitive aspects on which more and less proficient students differ, the features of assessment situations in which performances of interest are likely to be elicited, and the kinds of scoring systems that link cognitive functioning to student performance. The eventual intent is to offer an empirically-based framework to complement and support current creative and psychometric approaches to alternative assessment development.

CHARACTERISTICS OF PROFICIENT PERFORMANCE

A significant part of current investigation in cognitive psychology concerns the ability of people to reason, understand, solve problems, and learn on the basis of these cognitive activities. Much of this research has used analysis of verbal protocols as the basis for describing cognitive processes and structures commensurate with more or less proficient performance in knowledge-rich domains. The knowledge structures and cognitive processes that distinguish expert and novice problem solvers on classroom and textbook problems in elementary physics (e.g., Chi, Bassock, Lewis, Reimann, & Glaser, 1989; Chi, Feltovich, & Glaser, 1981) have been identified, as have the mental models and strategies used by effective and ineffective learners in simulated science laboratories (e.g., Schauble, Glaser, Raghavan, & Reiner,

1991; Shute, Glaser, & Raghavan, 1989). These studies and others (cf. Chi, Glaser, & Farr, 1988) have led to the description of general dimensions of problem-solving performance along which more or less proficient individuals in a particular domain differ (e.g., Glaser, 1992; Glaser et al., 1987).

Proficient individuals possess subject matter knowledge that is connected, related, and cohesive. In contrast, less proficient individuals have isolated or loosely related pieces of information. These students may be able to respond correctly to simple recognition questions, but their knowledge is not structured enough for adequate reasoning or for making inferences. As students acquire knowledge or become proficient in a subject matter, the nature of their responses undergoes qualitative changes as a consequence of the increasing interconnectedness of their knowledge. Whereas less proficient students make simple assertions to justify their actions, more proficient students generate complex arguments to support their opinions or reasons for the actions they take in solving problems. Cohesiveness of knowledge is further displayed in student explanations: Proficient students offer coherent, elaborated explanations; less proficient students offer fragmented or disjointed explanations of concepts and how they are related.

Not only do proficient students have a strong knowledge base; their knowledge is linked to conditions of applicability and procedures for use. Consequently, when given a problem, proficient individuals qualitatively assess the nature of that problem and construct a mental model or internal representation from which they can make inferences and test out alternative explanations or solution paths (cf. Gentner & Stevens, 1983). This internal representation supports the formulation of a knowledge-based plan and underlies engagement in efficient and principled problem solving.

Less proficient individuals, in contrast, generate surface feature representations that constrain their ability to think through the problem. These students engage in a trial-and-error process of matching problem features to school-learned examples. This problem-solving strategy proves ineffective when the surface features of the posed problem (e.g., ball and string) do not match the surface features of the examples studied in school (e.g., pendulum). In other words, these students, unlike their more proficient peers, are not aware of the underlying principles (e.g., laws of motion) that

link problems which on the surface look quite different. In the end, they fail to recognize the applicability of their knowledge in new situations.

Accurate estimation of task demands and how these demands match one's capabilities and prior knowledge is also characteristic of proficient performance. When engaged in problem solution, proficient individuals employ self-regulatory or executive skills to monitor and control their performance, efficiently allocating resources, checking their thinking, and adapting their strategy as needed. Less proficient individuals, on the other hand, underestimate the task demands and overestimate their knowledge and skills. Consequently, they fail to efficiently deploy resources to ensure task completion. Further, once the problem solution is set in motion, these students see it through to the end without checking the legitimacy of their strategy or response. This failure to monitor exhibits itself in the form of contradictory or illogical statements that go unnoticed by the individual performing the task.

In summary, then, the cohesive knowledge structures of proficient students facilitate their ability to plan, reason, explain, draw inferences, systematically solve problems, and monitor their own performance. These dimensions or characteristics of proficient performance provide an initial framework for studying the cognitive demands of alternative assessments. Students who score high on the assessment tasks should exhibit some characteristics of proficient performance if the tasks tap higher order thinking skills. If on the other hand, the tasks are fancy ways of packaging rote recall of facts or procedures, then a link between score and level of reasoning and understanding will not be evident.

ALTERNATIVE ASSESSMENTS

The project has worked with alternative assessment programs in Connecticut and California. In particular, we focus on three assessments, each a unique example of the kinds of alternatives being proposed for use in large-scale testing programs—Exploratory Investigation, Conceptual Integration, and Component Identification. These assessments are highly innovative in nature, as a well-articulated framework for the development of alternative assessments does not exist. Rather, development has been guided, for the most part, by the informed, creative intuitions of the developers

(primarily teachers and other educators) in conjunction with traditional psychometric concerns for reliability and validity. The intent here is to further guide this development by pointing out aspects of existing assessments and their corresponding scoring systems that provide links to higher order thinking skills that current assessment practices are designed to measure.

Exploratory Investigation tasks were designed as part of the Connecticut Common Core of Learning Assessment Project aimed at aligning assessment with educational goals (Baron, 1991; Baron, Forgione, Rindone, Kruglanski, & Davey, 1989). Working with teachers and educators, the Connecticut project developed tasks to test reasoning and problem-solving skills of students currently enrolled in high school science classes (Baron, 1991). The intent was to develop tasks and scoring criteria that would provide information to: (a) students to monitor their own performance, (b) teachers to monitor instruction, and (c) policy makers to monitor science education in the state (Baron, Carolyn, Greig, & Lomask, 1992).

In general, each task consists of three or four parts, some requiring individual work and some, group activity. The first part introduces the task to 12th-grade students, asking them to make some observations, provide a written description of the problem, formulate an initial hypothesis, and suggest possible ways of investigating the problem. The second part has groups of three or four students pool the observations and ideas of their group members and come to some consensus as to the salient variables in this particular task. As a group, students then generate a hypothesis, design and conduct experiments to test that hypothesis, document the tests and observations, and provide written conclusions based on their experiments. The last part of the task is answered individually and consists of a set of follow-up questions related to the task such as analyzing and critiquing a given set of data collected by an imaginary group on the same task.

Conceptual Integration tasks were also developed as part of the Connecticut Common Core of Learning Assessment Project (Baron et al., 1992; Lomask, Baron, Greig, & Harrison, 1992). These tasks were designed to assess students' understanding of science concepts regardless of their previous science experience or their current science enrollment. Like the Exploratory Investigation tasks, the information gained from an evaluation of task performance is intended to inform students, teachers, and policy makers.

Each task asks 11th-grade students to respond, individually, to several open-ended statements or interpret a science passage in free format about central topics in biology, chemistry, physics, and Earth science. Although the conceptual integration tasks are not performance tasks per se, they are open-ended and the student must decide which concepts are related and how they are related. Given our current understanding that organization and structure of knowledge are key distinguishing aspects of more or less proficient performance (e.g., Chi et al. 1981), conceptual integration tasks provide opportunities to examine this link in the context of student explanations of science concepts.

A *Component Identification* task was developed as part of a large research study examining alternative assessments of science achievement. The study, a collaborative effort among researchers from the University of California, Santa Barbara, California Institute of Technology scientists, and elementary school teachers in the Pasadena Unified School District, investigated the development and technical evaluation of hands-on performance assessments and more cost-effective alternatives such as laboratory notebooks, computer simulations, and paper-and-pencil measures (see Shavelson, Baxter, & Pine, 1991 for details). To date, the tasks developed as part of this study have undergone extensive psychometric analysis, and on that basis can be distinguished from the Exploratory Investigation and Conceptual Integration tasks; the latter are in a more formative stage of development.

The Component Identification task, one of three hands-on assessments developed in the Shavelson et al. (1991) study, asks 5th-grade students to apply their knowledge and skills in a novel problem-solving situation. Students are provided with equipment and asked to conduct tests to identify an unknown entity. To solve the problem students engage in a cyclical reasoning process of hypothesize, test, and refine. Because evaluation of student performance is used to monitor and inform instructional practice in a large urban school district in California, the tasks are administered immediately following an 8-week, hands-on, science instructional unit.

ANALYSIS

The assessments selected for use in this study vary with respect to the nature of the task (Exploratory Investigation, Conceptual Integration, or Component Identification), purpose of the assessment (student achievement, policy information, or instructional monitoring), stage of development (initial trials or following formal technical evaluation), grade level (12th, 11th, or 5th), and the assumptions of prior knowledge (little, general but unknown, or specific instructional experience). Given this diversity of tasks, student protocols were collected and appropriate analyses were performed to elucidate the cognitive activity underlying performance. Analyses of the detailed verbal protocols of students' performances in each of the assessment situations were guided by general dimensions of problem solving on which more or less proficient students differ. These dimensions are viewed as a general framework to focus the analysis and not a rigid conceptualization of the necessary requirements for alternative assessments. Some of the assessment tasks will readily lend themselves to certain dimensions (e.g., explanation) and not others (e.g., self monitoring).

Protocol analysis in conjunction with observations of students' performances and an examination of students' answer booklets and scoring criteria provide an empirical basis for linking performance scores with level and kind of reasoning and understanding. In taking this approach, it was reasoned that students who score high on the assessment task should exhibit some characteristics of proficient performance (ability to plan, reason, explain, draw inferences, systematically solve problems and monitor their own performance) if the task requires students to engage in higher order thinking skills. Further, the performance scores should differentiate among students with varying levels of proficiency. For example, students who provide fragmented explanations (low scores) should be distinguished from students who provide complete, coherent explanations (high scores). In the analysis then, two aspects were examined: (a) the task for opportunities to engage in higher order thinking, and (b) the scoring system for its ability to capture and reflect differential performance.

Exploratory Investigation

The Exploratory Investigation task examined in this study, "Exploring the Maplecopter," requires students to apply/draw on their knowledge of physics concepts of force and motion to reason about an everyday phenomenon—the flight of the maple seed. Students are required to hypothesize, predict, explain, and draw inferences. Further, the task does not have a clean, simple solution. Rather, students must rely on controlled experimentation and model-based reasoning to help them identify the causal variables involved in order to produce a convincing explanation of the "flight" of the maple seed.

Task

Exploring the Maplecopter consists of five scorable components: observation, experimentation, explanation of the motion of the maplecopter, advantages/disadvantages of models, and critique of sample experimental report (see Figure 1). During one class period, on each of 4 consecutive days, students complete all components of the task individually with the exception of the experimentation, which is conducted in groups. On Day 1, students study the motion of maple seeds (observation). On Days 2 and 3, students work in groups of three or four to design experiments to explain the spinning flight patterns of the maple seed. To encourage students to use models in their experimentation, directions for constructing a paper model of a helicopter are given to them. On the 4th day, students work individually to: (a) explain the motion of the maple seed, (b) reflect on the advantages and disadvantages of using models to study the motion of the maple seed, and (c) critique an experimental report, commenting on both the design and the validity of the conclusions.

Scoring

As part of the pilot assessment project, a scoring procedure was developed for teachers to use to evaluate student performance on the Exploratory Investigation task (see Figure 2). Students' written responses are read and scores are assigned for observation, explanation of the flight of the maple seed, reflection on the advantages and disadvantages of models, critique of sample report, and drawing conclusions from sample report. For each of these components of the task (observation, explanation, reflection, and application)

Part I: Getting Started by Yourself
Throw a winged maple seed up in the air or drop it from your hand. Observe how it "floats" down to the floor. Describe as many aspects of the motion of the pod as you can (you may add a diagram if you wish).

1. Record all observations that you have made. Do not explain the winged maple seed's motion at this time.
2. Try to explain how and why the winged maple seed falls as it does.

Part II: Group Work
Discuss the motion of the winged maple seed with the members of your group.

1. Write a complete description of the motion, using the observations of the entire group. (You may add a diagram if you wish.)
2. Write down all the factors that your group thinks might affect the motion of the winged maple seed.
3. Design a series of experiments to test the effects of each of these factors. Identify which of these experiments you could actually carry out.

Part III: Finishing by Yourself
Suppose you want to explain the motion of a winged maple seed to a friend who has not yet studied high school physics.

1. Write an explanation that is clear enough to enable your friend to understand the factors and forces which influence the motion of the winged maple seed. Specify the aspects about which you are more certain and those about which you are more unsure.

In this activity you used simplified models to help explain a more complicated phenomenon.

2. Explain all of the possible advantages and disadvantages of using models in studying the motion of a winged maple seed. Include specific examples from the models your group used.

Given a set of data generated by a group of students working on the maplecopter task. Read the report and answer the questions.

3a. Discuss the information given and how it is organized. Do you think it is complete enough for you to replicate the experiments? If not, what else do you need to know?

3b. Can any valid conclusions be made regarding variables studied in this experiment? If so, explain fully what they are.

Figure 1. Exploratory Investigation task, Exploring the Maplecopter (adapted from Baron, Carlyon, Greig, & Lomask, 1992).

Part I: Getting Started by Yourself	Excellent	Good	Needs Improvement	Unacceptable
1. Make observations	6 or more	4 - 5	2 - 3	0 - 1
1. Two phases: freefall & spinning 2. Velocity of freefall phase is greater 3. Tilted with seed lower 4. Rigid edge of the wing is leading edge 5. Spins around axis in seed 6. Spins either side facing up 7. spins either clockwise or counterclockwise 8. Motion different with different starting positions Others:				

Part III: Finishing by Yourself	E	G	NI	U
1. Explain the motion Holistic judgment based on the following: 1. Reference to or consistency with conclusions from experiments. 2. Inclusion of the forces and factors studied. 3. Explanation of physics is clear and appropriate to specified audience. 4. Lack of misconceptions.				
2. Explain the use of models.	4	3	2	0 - 1
Explanation should be based on following criteria: Advantages: 1. Materials are cheaper or more readily available/nondestructive of original. 2. Easier to control and manipulate variables/uniformity of models. Disadvantages: 1. Parameters of model are not the same as the "maplecopter" (i.e. shape, materials, etc.). 2. Uncertainty about the generalizablity of results from model to original.				
3.a. Critique information given in a research report.	5	4	3	0 - 2
The following should be included: 1. No definition of dependent variable. 2. No description of method. 3. Poor description of independent variables. 4. No description of model. 5. Poor organization of data.				
3.b. Critique conclusions made in research report.	4	2 - 3	1	0
Tentative conclusions can be made about the effects of the following: 1.a. Length of wing. 1.b. Added mass. 1.c. The stiffness of wing. 2. Conclusions are tentative due to uncertainty about accuracy of measurements.				

Figure 2. Scoring criteria for the Exploratory Investigation task, Exploring the Maplecopter (Connecticut Common Core of Learning Assessment Project, 1991).

BEST COPY AVAILABLE

teachers examine student responses for several critical aspects of task performance. On the basis of the match between student response and score criteria, individual student performance is described with respect to one of four levels—Excellent, Good, Needs Improvement, or Unacceptable.

Data Collection

Two interviewers visited a high school in Connecticut while the maplecopter task was being administered to several sections of seniors by their respective classroom teachers. The purpose of the interviews was to gather confirming or collateral evidence of the nature and extent of the cognitive activity required for successful task performance. To this end, eight students in a general physics class were interviewed and audiotaped each of the 4 days they worked on the maplecopter task. The interview, guided by the nature of students' answers to the part of the task they had completed that day, provided an opportunity for students to explain and/or elaborate on their written responses. Students were also asked to list all the physics concepts learned in school that they thought were relevant to the maplecopter task and explain how or in what way they thought these concepts were related.

Students' written responses and the teacher's evaluation of those responses (using the scoring criteria developed in conjunction with the Exploratory Investigation task) were collected for the eight general physics students. In addition, written responses of six physics students in an Advanced Placement (AP) class and the teacher's evaluation of these students were also collected. Due to competing demands, these six students were not interviewed.

Results

We examined the relationship between student scores and level and kind of cognitive performance elicited for each of the four components of the assessment task students completed individually. For ease of presentation we have labeled each of these components as follows: Observation, Explanation, Reflection, and Application. Student scores were assigned by the classroom teacher according to the directions and specifications for judging performance set forth by the test developers (see Figure 2).

For each of the components of the maplecopter task we present: (a) an overview of the expected cognitive activity defined by the nature of the task, (b) a description of the kind and extent of cognitive activity students engage in as determined by interviews, and (c) the results of an examination of the degree to which evaluations of students' performances (scores) are linked to characteristics or dimensions of proficient performance.

Observation. Part I of the Exploratory Investigation task asks students to make observations of the maple seed and to offer an explanation of their observations (see Explanation below). Scientific observation is an important first step toward successful experimentation and might be considered a critical part of any open-ended performance assessment of this type. "Learning to observe is not just a matter of acuity of vision," but "involves making decisions as to which features are relevant and which can be ignored" in the given context (Millar & Driver, 1987, p. 47). Adequate decision making, in this context, is dependent on students' reasoning about the number and quality of the observations they make and monitoring their thinking in an effort to "describe as many aspects of the motion of the maple seed" as they can. This observational process begins with a hypothesis about an aspect thought to be important (e.g., size of the maple seed), followed by systematic testing of that hypothesis (e.g., observing the motion of seeds of different sizes), and ends when conclusions are drawn (e.g., the larger seeds fall slower). This sequence of hypothesize, test, draw conclusions is repeated until students decide the important aspects have been noted and the observation, therefore, is complete.

In this task, some aspects of the flight of the maple seed are more readily observable than others. For example, even the casual observer would notice that there are two phases—an initial freefall followed by a spinning stage. Other observations require a more focused and informed look (e.g., rigid edge of the wing is the leading edge). Nevertheless, all aspects are given equal weight in the scoring criteria. It appears that the primary intent was to have students describe as many aspects of the flight of the maple seed as they could, based on their observations. Student performance is scored solely on the number of relevant aspects observed without an indication of how the observations were made or any justification for including particular aspects as relevant or influential with respect to the flight of the maple seed.

An examination of students' written responses indicates that all students, AP and general physics, observed the two phases of freefall and spinning. The exceptions were two students who observed aspects of the spinning stage but not the freefall stage "since that is not what the experiment is about I just didn't worry about it." Moreover, students described, at most, four of the eight aspects listed in the scoring criteria (see Figure 3). Furthermore, the particular aspects described varied within and across the two physics classes (general and AP) with the latter providing slightly more observations, on average, (three) than the former (two).

Although, on average, the general physics students mentioned two aspects, different students mentioned different aspects. No aspect was predominant across the responses except the two phases of freefall and spinning. Aspects such as the "maple seed spins either side facing up" or "the maple seed spins either clockwise or counterclockwise" were not mentioned by any of the students. Indeed these two aspects are related; direction of spin is dependent on which side of the seed is facing up. The seed always spins with the soft edge trailing the rigid edge.

There was slightly more consistency among the AP students in the particular aspects observed. All students mentioned the two phases and that the seed "spins around axis." Other aspects seemed less relevant to these students such as the seed spins with either side facing up or that the motion is different with different starting positions (aspects five and eight, respectively, of the scoring criteria; see Figure 3).

During the interviews, the eight general physics students were asked how many observations they had made or how many observational trials they had conducted (i.e., number of times they dropped the maple seed). Although the responses ranged from "don't know" to 2 to 10 or 20, the majority of the students (six of eight) conducted less than 10 observational trials. Further, there was no systematic relationship between the number of observational trials conducted and the number of aspects reported as relevant to the motion of the maple seed. A student who only mentioned the two phases of the maple seed flight (freefall and spinning) had dropped the maple seed 10 times while a student who had noticed three different aspects (two phases, velocity of freefall greater, and spins tilted with seed lower) conducted only 2 observational trials.

Part I: Getting Started by Yourself		Excellent	Good	Needs Improvement	Unacceptable
1. Make observations		6 or more	4 - 5	2 - 3	0 - 1
1. Two phases: freefall & spinning	5. Spins around axis in seed				
2. Velocity of freefall phase is greater	6. Spins either side facing up				
3. Tilted with seed lower	7. spins either clockwise or counterclockwise				
4. Rigid edge of the wing is leading edge	8. Motion different with different starting positions				
Others:					

Figure 3. Scoring criteria for the observation component of the Exploratory Investigation task, Exploring the Maplecopter (Connecticut Common Core of Learning Assessment Project, 1991).

BEST COPY AVAILABLE

To summarize, students, regardless of current physics enrollment, observed, at most, four of the eight aspects of the flight of the maple seed listed on the score form (see Figure 3). Evidence from the interviews indicates that few students engaged in repeated, systematic observational trials necessary for reliable identification of key aspects involved in the motion of the maple seed. Students' scores for this component of the task were correspondingly low, ranging from 1 (unacceptable) to 3 (needs improvement). Proficient students (those who know what reliable observation entails) cannot be distinguished from less proficient students on the basis of their score.

More specifically, the scoring system does not recognize or display proficient performance as an iterative sequence of systematic observation—observe, test, conclude, observe—in such a way that students' performances are evaluated with respect to the degree to which they engage in this process. Rather, the scoring procedure, as currently conceived, is based on students generating a list of relevant aspects. Our interviews suggest that students, regardless of their score (1 or 3), may or may not have engaged in systematic observational trials. For this task, then, scores do not differentiate more from less proficient performance.

Explanation. As part of the Exploratory Investigation task, students were asked to provide two explanations: (a) Explain how and why the winged maple seed falls as it does based on the observations you have made; (b) Explain the motion of the winged maple seed to a friend who has not yet studied high school physics based on the experiments you conducted. Evaluation of student explanations provides information on the cohesiveness of student knowledge structures, a key distinguishing feature of proficient performance (cf. Glaser et al., 1987). Less proficient students (those with less structured knowledge) provide fragmented or disjointed explanations. In contrast, proficient students (those with well-structured knowledge) provide more complete, coherent explanations of the important concepts and the relations among them. If performance scores for this component of the Exploratory Investigation task are linked to the underlying structure of knowledge, then students who score high (i.e., proficient performance) should (a) know more task-related concepts, and (b) reflect that understanding in a coherent written explanation of the relations among those concepts.

The pre-experimental explanations based on students' initial observations of the maple seed were not targeted for scoring by the assessment developers (see Figure 2). In our analysis of the post-experimental explanations, we examined the relationship among students' knowledge of task-related concepts gained from interviews, their performance scores assigned by the teacher according to the score criteria, and the quality of their post-experimental explanations.

First, the number of concepts students identified in their interview was compared with the quality of their explanations. Proficient students would be expected to know more concepts and be able to relate those concepts in a coherent explanation. The quality of students' post-experimental explanations were evaluated with respect to one of five levels as follows: Level 1 (Nonexplanation), student mentions one or two factors (e.g., location of center of mass, blade size and shape, overall mass) or forces (gravity, air resistance); Level 2 (Fragmented), student mentions one or more forces and factors but with little elaboration of their impact on the flight of the maple seed; Level 3 (Partial), student gives an elaborated, coherent description of a subset of forces and factors; some factors or forces were not mentioned and/or there was some evidence of major misconceptions (e.g., terminal velocity causes the seed to start spinning); Level 4 (Good), student gives an elaborated, coherent description of most of the factors and forces, and makes reference to his or her own experiments; Level 5 (Ideal), student gives an elaborated, misconception-free description of all the factors and forces and how they impact each of the two phases with reference to his or her own experiments.

Results indicate that although one-half of the general physics students who were interviewed identified several of the relevant concepts (center of mass, air resistance, gravity and terminal velocity), they were unable to explain the relations among these concepts within the context of the flight of the maple seed. Three of the eight students provided a fragmented explanation and the remaining five students provided a partial explanation with respect to some of the relevant forces or factors. Four of the six AP physics students provided good explanations, albeit with some misconceptions (see Figure 4). Regardless of current enrollment, students in this sample did not provide nonexplanations or ideal explanations.

Level 2 (Fragmented)

"Gravity usually makes an object fall at about 9.8 m/s^2. Because the maple leaf and maplecopters have so much air resistance they didn't fall at even close to that. The center of mass was at the base, and because of this the maplecopter spun around rather than falling straight down. If the mass was in the middle rather than at one end it would have fallen straight down. The smaller the leaf, the faster it would spin. The way it fell depended on its mass, its size, and its wings."

Level 3 (Partial)

"The winged maple copter dose things when it falls that you do everyday. First, it is similar to when you fall from a tree. What brings you back to the ground is the same thing that brings the maple seed to the ground. It is called gravity. Secondly, the maple copter is affected by the force that you feel when you stick your head out of a car window, when your mother tells you not to. You feel a wind that wants to slow down your head. This is the same force that keeps the maple copter from falling quickly. It is called air resistance. Lastly, is centripetal force. Just as it affects you on the swing set, so does it affect the maplecopter. Just a[s] the ropes of the swing stay taunt when you are up in the air and keep you in an arck. The seed keeps the wing from going straight, and thus the wing holds the seed up."

Level 4 (Good)

"There are several variables that affect the motion of the winged maple seed; the curve, the weight of the seed with respect to the wing, the weight of the "hard edge," the surface area/air resistance. By looking at these aspects one can gain a simple understanding of why the maple seed spins as it does. When the seed is initially dropped, it can be seen that the heavier end, the seed, leads the way down. This is understandable as it weighs more than the rest of the wing. So why doesn't the maple seed continue to fall in this manner? Because of the air resistance that it encounters. As the light wing of the maple seed accelerates, the air resistance continues to build. This is where the curved body becomes important. As the air begins to build under the wing, it searches for an escape. The curve is concave up to the side without the "hard edge" and so the air escapes to this side. Subsequently, a force is produced that results in a spinning motion in the direction of the "hard edge." In coming to this conclusion, many experiments were conducted. Adding more mass to the seed will increase the time required to reach terminal velocity (spin begins). This is because if the weight of the seed is increased, the maple seed will drop faster and so it will take longer for an adequate air resistance to build up to produce a spin. Air resistance is also affected by the surface area of the wing. An increase in surface area will result in a shorter time period between dropping and the starting of spinning. It would have been ideal to have been able to test the flight of the maple seed in a sealed vacuum. It is believed that the maple seed would have fallen seed first the whole way down as air resistance would have been removed."

Figure 4. Examples of student explanations.

Next, we compared the performance scores based on the scoring criteria with the quality of students' explanations. We expected students with the

highest scores (most proficient) to provide quality explanations (complete and coherent). Scores were obtained from the teachers who administered the assessment to their respective classes. In assigning a score, teachers read the students' written responses and holistically evaluated the explanations of the motion of the maple seed with respect to four criteria (see Figure 5). Given the complexity of the task, making a holistic evaluation of students' explanations without student exemplars representative of each of the score levels is not a trivial matter. Rather it requires in-depth knowledge of the maple seed and the forces and factors that impact its "flight." Indeed, understanding how and why the maple seed falls as it does has drawn attention from a broad spectrum of researchers including biologists, biophysicists, and aerospace engineers (e.g., Green, 1980; Seter & Rosen, 1992; Ward-Smith, 1984).

Given the complexity of the task and the nonspecific nature of the scoring criteria used to judge performance on the task, it is not surprising that scores teachers assigned using the scoring system differed slightly from a qualitative evaluation of students' explanations conducted as part of this study (see Table 1). In general, for those students who scored high, there was considerable consistency between evaluation based on the score criteria and evaluation of the quality of students' explanations. There was less consistency for students who scored low. Some students who provided a Level 2 explanation (fragmented) were judged on the basis of the score criteria as "needs improvement." Others who provided a Level 3 explanation (partial) were judged as "needs improvement." Clearly the evaluation is correct; improvement is needed. Nevertheless, important distinctions are blurred. If the tasks are to provide meaningful, instructionally relevant feedback to students and teachers about task performance, clear distinctions are necessary.

To summarize, the explanation component of the Exploratory Investigation task provides an opportunity for students to demonstrate their understanding of concepts studied in physics class in the context of explaining a real world phenomenon—the flight of the maple seed. In general, AP students provided more coherent explanations than students in regular physics classes. While students in the general physics class could list several of the relevant physics concepts, they were unable to provide more than a partial explanation of the relations among those concepts in the context of the

Part III: Finishing by Yourself	Excellent	Good	Needs Improvement	Unacceptable
1. Explain the motion	E	G	NI	U
Holistic judgment based on the following: 1. Reference to or consistency with conclusions from experiments. 2. Inclusion of the forces and factors studied. 3. Explanation of physics is clear and appropriate to specified audience. 4. Lack of misconceptions.				

Figure 5. Scoring criteria for the explanation component of the Exploratory Investigation task, Exploring the Maplecopter (Connecticut Common Core of Learning Assessment Project, 1991).

Table 1

Relationship Between Performance Score and Quality of Written Explanation

Science class/ID	Performance score[a]	Quality of explanation[b]
AP physics		
11	Good	Good
12	Good/Excellent	Good
13	Good	Good
14	Needs Improvement/Good	Fragmented
15	Good	Good
16	Needs Improvement	Fragmented
General physics		
21	Needs Improvement	Partial
22	Good/Excellent	Partial
23	Unacceptable	Fragmented
24	Needs Improvement	Partial
25	Needs Improvement	Fragmented
26	Needs Improvement/Good	Partial
27	Needs Improvement	Fragmented
28	Needs Improvement	Partial

[a]Score assigned by the teacher using the scoring criteria developed in conjunction with the assessment task (see Figure 5).

[b]Qualitative evaluation of quality of students' explanations (see Figure 4).

maplecopter task. The scoring criteria reflect this lack of knowledge and understanding about the forces and factors influencing the flight of the maple seed by summarizing the performance generally as "needs improvement." In doing so, instructionally important distinctions among students' performances are not apparent. Providing guidance for scoring in the form of student exemplars will help scorers to distinguish proficient from less proficient performance with respect to instructionally relevant differences among students' performances.

Reflection. The flight of the maple seed "represents a delicate equilibrium between gravity, inertia, and aerodynamic effects. Therefore, in order to

analyze this phenomenon, an accurate detailed model is necessary" (Seter & Rosen, 1992, p. 196). The third component of the task asks the general physics students to "explain the advantages and disadvantages of using models in studying the motion of a winged maple seed. Include specific examples from the models your group used." For the AP physics students the question was worded slightly differently. "Describe the advantages and disadvantages of using your model as a model of a winged maple seed." (Recall, this task is undergoing development and changes are being made based on teacher feedback.) Despite these minor differences in wording, both these questions ask students to reflect on their experiences with models for studying the maple seed and on that basis explain the advantages and disadvantages of doing so.

Reasoning about the effectiveness of using models to understand a complex, naturally occurring phenomenon—flight of the maple seed—requires students to reflect on their own experimentation with models, noting the strengths and limitations of model-based reasoning. We might expect proficient students, then, to include in their response a list of advantages and disadvantages with a justification for each based on their own experiments.

An examination of students' written responses indicates that three students listed two advantages and two disadvantages, but only two general physics students made reference to the models they had constructed for their experiments. All of these students received the same score (i.e., excellent) because students' written explanations were judged on the basis of a match with a generic, context-free list of two advantages and two disadvantages of using models (see Figure 6). No mention was made in the score criteria of "reference to own experiments."

Four of the AP students and one general physics student scored 3 of the 4 possible points, but none of these students made reference to their experiments with models. Recall AP students were not specifically asked to give examples based on their experiments but rather to reflect on their experiences with their model—"Describe the advantages and disadvantages of using your model as a model of a winged maple seed."

For this component of the task, students' scores were not linked to level of reasoning and understanding. This may, in part, be due to the mismatch between task requirements and criteria for judging student performance.

Part III: Finishing by Yourself	Excellent	Good	Needs Improvement	Unacceptable
2. Explain the use of models.	4	3	2	0 - 1
Explanation should be based on following criteria: Advantages: 1. Materials are cheaper or more readily available/nondestructive of original. 2. Easier to control and manipulate variables/uniformity of models. Disadvantages: 1. Parameters of model are not the same as the "maplecopter" (i.e. shape, materials, etc.). 2. Uncertainty about the generalizablity of results from model to original.				

Figure 6. Scoring criteria for the reflection component of the Exploratory Investigation task, Exploring the Maplecopter (Connecticut Common Core of Learning Assessment Project, 1991).

Students are asked to provide an explanation of the advantages and disadvantages of using models; they are given credit for providing a list with no accompanying rationale. While even the least proficient students may be able to generate a list, only the most proficient can explain the advantages and disadvantages with respect to their own experimentation with models. If students' performances are not to be over- or underestimated, a link between performance score and level of reasoning and understanding is critical.

Application. The fourth and final component of the task presents students with a sample group report. Students are asked to read the report generated by a group of students who had worked on the maplecopter task and comment on: (a) organization of the report, and (b) the validity of the conclusions drawn (see Figure 7). To adequately critique someone else's experiment, students draw on their personal knowledge and understanding of experimental procedures, in general, and their task-specific conceptual understanding of the factors and forces involved in the flight of the maple seed. Using this knowledge and understanding as the criteria, students engage in a mapping process—mapping the group report onto their own conception of an experiment and, in particular, the maplecopter experiment—checking for matches and mismatches.

In our analysis we examined the consistency between students' understanding of experiments gained from their interviews and their performance score. Performance scores were assigned by teachers on the basis of a match between students' written responses and score criteria (see Figure 8). Separate performance scores were given for each part of the application component (organization and validity of conclusions). We would expect proficient students (high scores) to have a better understanding of the experimental process.

As was the case with the observation and explanation components of the maplecopter task, there was a considerable performance difference between AP and general physics students, with the former outscoring, on average, the latter. Moreover, the results were similar for questions 3a (critique of design) and 3b (validity of conclusions). In critiquing the experiment, all of the eight general physics students provided inadequate responses. The best answers from this group were from two students who identified poor organization of

THE GROUP REPORT

We tested paper helicopters to see if different lengths (3), stiffness (1) and weight (3) would affect the helicopter.

We used:

1) 3 cm wing length, stiff (4 paper clips)
2) 6 cm wing length, stiff (4 paper clips)
3) 0 cm wing length, stiff (4 paper clips)
4) 6 cm wings, flexible (4 paper clips)
5) 1/2 way cut through 10 cm wings stiff (4 paper clips)
6) 3/4 way cut through 10 cm wings stiff (4 paper clips)
7) 3 paper clips on 10 cm wings stiff (3 paper clips)
8) 5 paper clips on 10 cm wings stiff (5 paper clips)

Data:

Type	Average Time
1) 3 cm-s	.49 sec.
2) 6 cm-s	.66 sec.
3) 10 cm-s	1.29 sec.
4) flexible 6 cm	.77 sec.
5) 1/2 cut-s	1.07 sec.
6) 3/4 cut-s	.97 sec.
7) 3 pc-s	1.15 sec.
8) 5 pc-s	1.21 sec.

Our data confirmed our belief that wing length, stiffness, and weight would affect the helicopter. The results turned out as expected.

3a. Discuss the information given in the report and how it is organized. Do you think it is complete enough for you to replicate the experiments? If not, what else do you need to know?

3b. Can any valid conclusions be made regarding the variables studied in this experiment? If so, explain fully what they are.

Figure 7. Critique of group report (adapted from Connecticut Common Core of Learning Assessment Project, 1991).

data and no description of method as deficiencies in the report. Likewise their responses to the question addressing validity of the conclusions received scores of zero or 1 with the exception of one student who received 4 points. When students were asked in their interview what makes a good experiment, the following response was typical. "I'm not . . . it's hard to say. It depends on what . . . I think it's the results you're getting from your experiment is the judge of a good experiment."

Part III: Finishing by Yourself	Excellent	Good	Needs Improvement	Unacceptable
3.a. Critique information given in a research report.	5	4	3	0 - 2
The following should be included: 1. No definition of dependent variable. 2. No description of method. 3. Poor description of independent variables. 4. No description of model. 5. Poor organization of data.				
3.b. Critique conclusions made in research report.	4	2 - 3	1	0
Tentative conclusions can be made about the effects of the following: 1.a. Length of wing. 1.b. Added mass. 1.c. The stiffness of wing. 2. Conclusions are tentative due to uncertainty about accuracy of measurements.				

Figure 8. Scoring criteria for the application component of the Exploratory Investigation task, Exploring the Maplecopter (Connecticut Common Core of Learning Assessment Project, 1991).

In contrast, all of the AP students received a score of 4 or 5 for their critique of the group report and for their judgements of the validity of the conclusions. This suggests that these students, unlike their general physics peers, have a clear conception of what constitutes a good experiment and can use this understanding to identify weaknesses in the experimental reports of others.

Summary

An Exploratory Investigation task, of which "Exploring the Maplecopter" is an example, is clearly consistent with the goals of science education. Drawing on their knowledge of physics, students are asked to engage in model-based reasoning to investigate and describe an everyday phenomenon—flight of the maple seed. To understand and explain relations among causes and effects requires sustained and broad exploration, beginning with observation, followed by experimentation, and interpretation of findings.

In the maplecopter task, students appeared to initially interpret the task as one of engineering as evidenced by their attempts to build a model that replicated the maple seed's motion, or a model that stayed aloft for the longest time. In contrast, the purpose behind using models is to allow students to understand experimentation by manipulating and controlling various aspects of the maple seed that may influence the "flight." Students' interpretation of the task and/or lack of experience with an exploratory investigation of this type may account for the difficulty they had in getting started.

Nevertheless, after engaging in observations of the important aspects of the maple seed, all the students built models and conducted experiments. Based on their experiments with these models, they generated an explanation of the flight of the maple seed. They then reflected on the advantages and disadvantages of using models, and finally they critiqued the design and validity of the conclusions drawn from a maplecopter experiment reported by another group of students. Each of these four components (observation, explanation, reflection and application) was scored on the basis of a match between students' written responses and scoring criteria.

Results of an analysis of student protocols in conjunction with an examination of student answer booklets and the scoring criteria indicate that

all components of the task clearly distinguished students on the basis of their current enrollment; AP physics students outperformed general physics students. Regardless of current enrollment, all students displayed relatively poor observation skills. Moreover, general physics students were aware of some of the relevant physics concepts related to this task but had difficulty relating those concepts in a coherent explanation of the target problem—the motion of the maple seed. The least discriminatory component was the advantages and disadvantages of using models. This may be due to the discrepancy between what students were asked to do and the criteria listed on the score form. Students were asked to provide examples, but the criteria for scoring consisted of a list of two advantages and two disadvantages. Although all students could list one or more advantages/disadvantages, it is not clear whether all students could reason about those advantages and/or disadvantages with respect to their own models and experiments. The critique of the experiment, on the other hand, is related to students' understanding of experimental procedures. Students with little understanding of experimental procedures in general, or the specifics of the maplecopter experiment, had difficulty in adequately critiquing the experiments of others. Finally, performance on one component of the Exploratory Investigation task was reflective of performance on the other components. In general, those students who made the most observations also tended to give better explanations and more adequate experimental critiques.

The Exploratory Investigation task, by virtue of its open-ended structure, provides students the opportunity to display their level of knowledge and understanding. Students were asked to observe, explain, reflect and apply their knowledge to critique the work of others. Results of our analysis indicate that, for this task, quality explanations of the flight of the maple seed, critiques of the experiments of others, and consistent performance across all components of the task are evidence of cohesive knowledge structures prerequisite for proficient performance.

The task is rich with opportunities for students to engage in higher order thinking, but the scoring system fails to capture the essence of proficient student performance. First, students received a score of 1 to 4 based on the number of aspects observed and not on the process in which they were engaged. Indeed, student interviews suggest that very few students have a

clear conception of what reliable observation entails. Second, students received feedback on their explanations in the form of "needs improvement" but without an accompanying example of a quality explanation. As such, important distinctions among students with varying levels of proficiency were blurred. Third, declarative knowledge of advantages/disadvantages of using models served as the score criteria for reflection on one's experiences with models. Even the least proficient students may be able to generate a list; however, only the most proficient can explain the advantages and disadvantages with respect to their own experimentation with models. Developing scoring systems that make apparent the characteristics of proficient performance are critical if these assessments are to accurately inform students, teachers and policymakers.

Conceptual Integration

The Conceptual Integration tasks ask students to provide an explanation of a major topic selected from one of several domains of science. Specific topics in biology, chemistry, physics and Earth science were chosen because they represent the kinds of knowledge that students are expected to have after high school, regardless of the particular science courses they have taken. As such, the topics chosen (e.g., photosynthesis) are a sample of those which students encounter in one of several courses (e.g., ecology, life science, biology). Moreover, students revisit these topics several times during their schooling, typically beginning in fourth or fifth grade. Most students, then, have had multiple opportunities to study each of the topics presented in this task.

In developing an explanation for each of the topics, students must make decisions about which concepts are important and how these concepts are related. Student explanations, then, serve as evidence of students' conceptual understanding. While a coherent and cohesive narrative is evidence of deep understanding of particular concepts, a set of unconnected statements is evidence of fragmented pieces of knowledge.

Task

Students are asked to respond in writing to each of the following: *Growing plants*—describe the types of energies and materials involved in the process of a growing plant and explain how these energies and materials are

related; *Digestion of a piece of bread*—describe the possible forms of energy and types of materials involved in the digestion of a piece of bread and explain fully how they are related; *Blood transfusions*—state what you would want blood to be checked for and explain why the blood should be checked for each of these if the blood is to be used for a transfusion.

Scoring

The Connecticut Common Core of Learning Assessment Project, drawing on the work of Novak and Gowin (1984), developed concept maps for scoring the Conceptual Integration tasks. For each of the tasks, these concept maps provide pictorial representations of core concepts and how they are interrelated. Concepts were considered core if teachers felt students should reasonably be expected to know them at this point in their schooling (Lomask et al., 1992). In other words, the concept maps focus on a subset of knowledge related to a particular topic and not all the possible concepts that could be mentioned (see Figure 9).

An expert's (teacher's) concept map serves as a "template" against which students' performances are evaluated. Students' explanations for each topic are read, and matches and mismatches with the expert concept map are noted. Teachers put a check on the expert map indicating a match with the students' explanations. On occasion students may mention concepts not considered core concepts, and these are added to the expert map. Further, students' explanations often reveal misconceptions that are also noted on the concept map. Thus the concept map provides a visual display of students' performances and how each corresponds to "expert" or expected performance.

Scoring focuses on two structural dimensions of the concept map— size and strength. Size is defined as the number of concepts included in students' explanations over the total number of core concepts in the expert concept map. Students are not given credit for mentioning relevant concepts in their explanations if those concepts are not displayed in the expert concept map. In other words, knowing more than is expected is not rewarded in this case (but see below).

Strength is defined as the number of valid connections in students' explanations over the number of possible connections. The number of possible

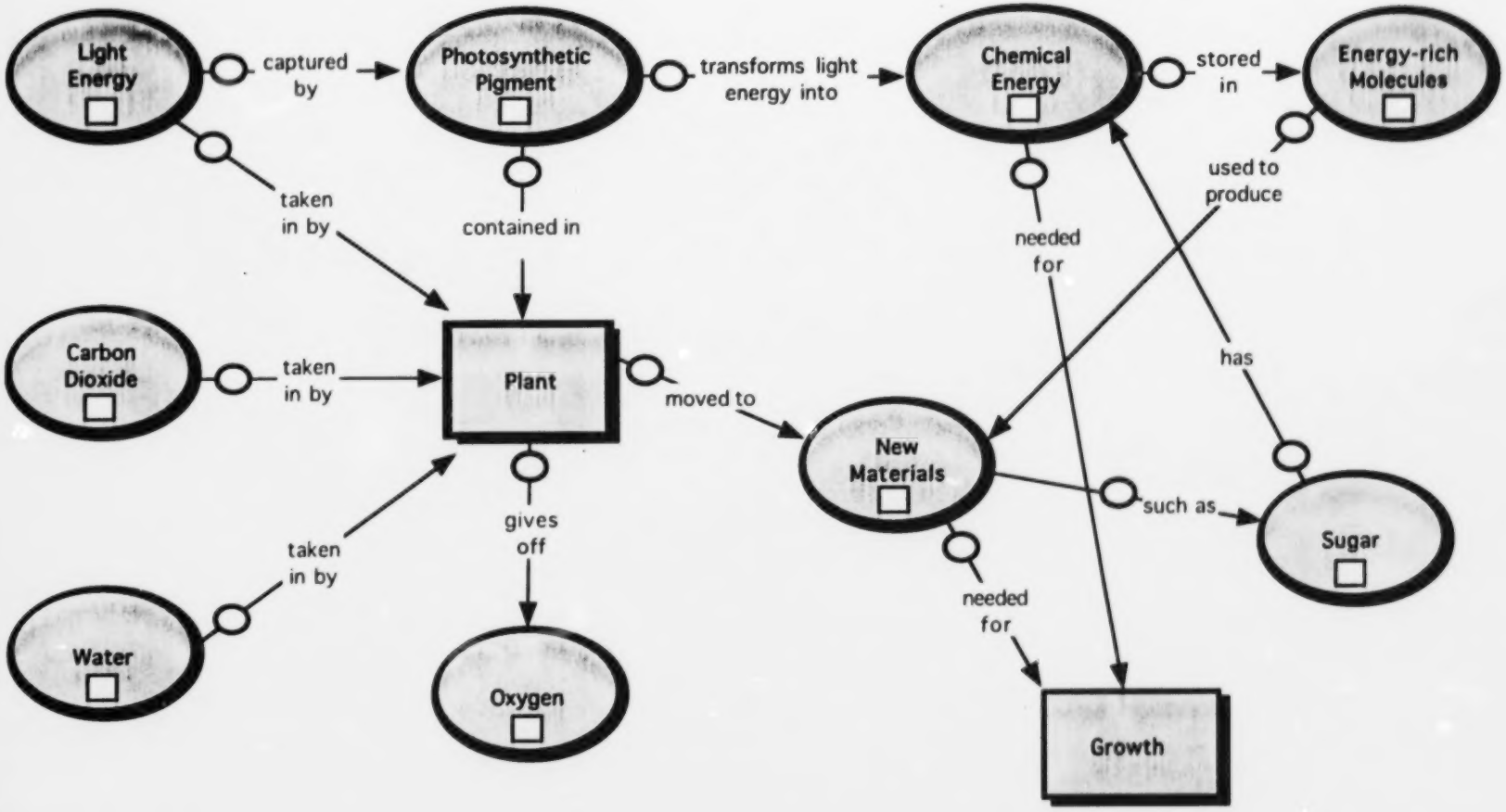


Figure 9. Concept map for scoring the Conceptual Integration task, Growing Plants (adapted from Lomask, Baron, Greig, & Harrison, 1992).

BEST COPY AVAILABLE

connections is determined by the number of core concepts mentioned. If students did not mention particular concepts, they would not be penalized for failing to provide information about the connections among those concepts. However, if students mention incorrect connections, they are penalized (strength score is reduced). Further, if students mention correct connections to concepts not mentioned in the expert map they are given credit (strength score is increased). The strength thus indicates if students know the "full story" at least about the concepts mentioned in their explanation.

Consider the Blood Transfusion question in working through an example of the scoring system. The question reads: "If you were to receive a blood transfusion, for what would you want the blood you were receiving to be checked? Explain fully why blood should be checked for these things." One student responded as follows:

> "Well, first of all I would want it checked to see if the blood was the same blood type as mine was. I would want it checked for diseases."

The teacher's evaluation of this student's response is shown in Figure 10. In looking at the concept map the student was credited for 2 (blood type and disease) of the 10 possible core concepts. Given that the student mentioned two concepts, the number of possible connections is also two. The student was credited for two connections (is checked for). For this student, the size score = 2/10 and the strength score = 2/2.

An overall score or evaluation of student performance is determined by combining size and strength scores. First, size and strength scores are assigned categorical labels reflective of the degree of overlap with the expert's explanation. Size is described with respect to five categories ranging from Irrelevant (student mentions less than 12% of the core concepts) to Complete (student mentions more than 82% of the core concepts). Strength is described with respect to three categories based on the proportion of necessary, accurate connections students make given the concepts they mentioned. The categories are Strong (67% or more of the possible connections given the concepts mentioned), Medium (33% to 67%) and Weak (less than 33%). Second, these categorical labels are displayed in a size-by-strength matrix (see Figure 11). Third, the overall score (where the size row crosses the strength column) is

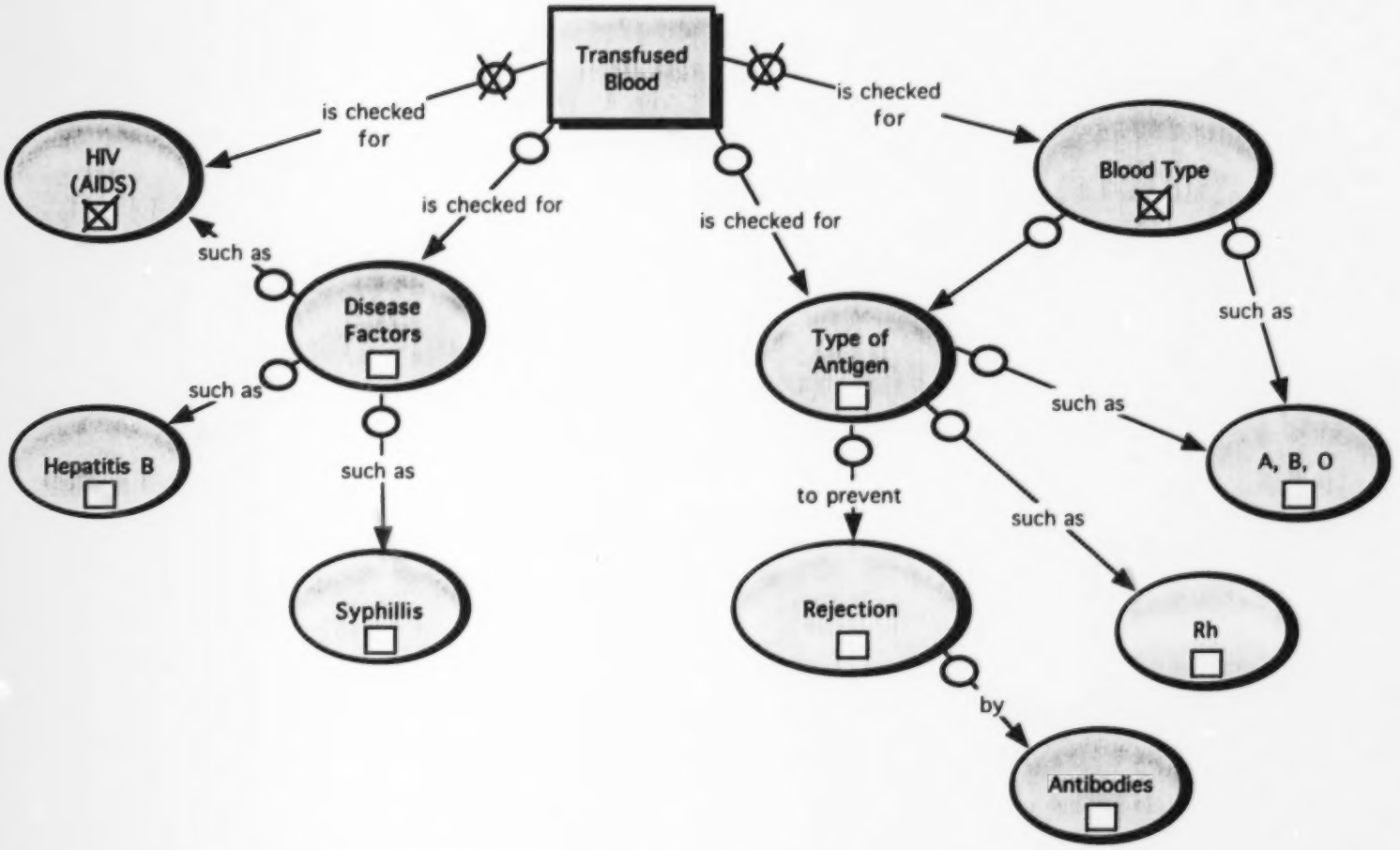


Figure 10. Example of a concept map with student response indicated.

BEST COPY AVAILABLE

Size[b]	Strength[a]		
	Strong > 67%	Medium 33% to 67%	Weak < 33%
Complete (>82%)	5	4	3
Substantial (58%–82%)	4	3	2
Partial (35%–58%)	3	2	1
Small (12%–35%)	2	1	1
Irrelevant (<12%)	1	1	1

[a] Strength = proportion of necessary, accurate connections between concepts mentioned in student explanation.

[b] Size = proportion of core concepts mentioned in student explanation.

Figure 11. Level of understanding for Conceptual Integration tasks (adapted from Lomask, Baron, Greig, & Harrison, 1992).

represented as a level of understanding on a scale from 1 to 5. A score of 5 (the highest level of understanding) indicates that students have a well-integrated conceptual knowledge of a particular topic (i.e., size is complete and strength is strong). Note that each level of understanding from 1 through 4 appears in the size-by-strength matrix more than once. For example an overall score of 4 might reflect size as complete or substantial with corresponding strength as medium or strong, respectively.

Referring to the student explanation of blood transfusion cited in Figure 10, size is 2/10 which is categorized as small because the student mentioned only 20% of the core concepts in the expert map. For strength, the student mentioned two out of the two possible connections or 100%. Strength is therefore categorized as strong (more than 67%). For this student (reading down the "strong" column and across the "small" row) the level of understanding is 2.

Data Collection

Twelve students with varying science backgrounds (n = 4 AP biology, n = 4 human biology, and n = 4 geology) were interviewed after they completed the Conceptual Integration tasks to elicit information in support of the cognitive validity of the assessment task. To this end students were asked to elaborate on the concepts mentioned in their written explanations. Further, they were asked questions about (a) their interpretation of each of the tasks (e.g., What do you think this question is asking?), (b) their understanding of particular concepts not mentioned in their written responses (e.g., Do you know what the Rh factor is?), and (c) their prior experiences with each of the topics (e.g., Have you studied this before?).

Along with the interview protocols, students' written explanations and teachers' evaluations of those explanations using the concept map scoring system were collected. In addition, we evaluated each explanation with respect to completeness and coherence, assigning scores from 1 to 5 as follows: Level 1 (Nonexplanation), student provides a list of core concepts or a list of statements; Level 2 (Fragmented), student provides an explanation which is incomplete (a subset of the core concepts mentioned) and lacks coherence with respect to the relationships among the concepts; Level 3 (Partial), student provides a coherent explanation of a subset of the core concepts; Level 4 (Good), student provides a coherent explanation of most of the core concepts. A few concepts may not be mentioned and/or a misconception may be evident; Level 5 (Ideal), student provides a complete, coherent explanation consistent with the expert concept map.

Results

In developing a response to the Conceptual Integration tasks, students reflect on what they know about the given topic, and decide which concepts are important and how they are related. If the concept maps "elucidate the conceptual structure of a student's scientific knowledge" (Lomask et al., 1992, p. 5), then level of understanding (i.e., concept map score) should correspond to the quality of student explanation. Our analysis, then, focused on the relationship between knowledge of task-related concepts as evidenced in the interviews, quality of students' written explanations (i.e., nonexplanation, fragmented, partial, good, or ideal), and level of understanding (1 to 5) as

determined by the scoring criteria developed by the Connecticut Common Core of Learning Assessment Project. We expected students who scored high to (a) know more task-related concepts as evidenced by their interviews, and (b) reflect that understanding in a coherent, written explanation of the relations among those concepts.

Digestion of Bread. Students were asked to "Describe the possible forms of energy and types of materials involved in the digestion of a piece of bread and explain fully how they are related." Interviews revealed that students viewed the task as having multiple possible interpretations. "Do you want me to go through the digestive system or do you want like cell respiration or did you want, you know, how the body uses the energy or what it's used for or . . ." In the end, students made a decision about the interpretation they would respond to, perhaps based on what they felt they knew best. "I don't remember like everything about the cell, you know, and all that."

Eight of the 12 students interpreted the question to mean "describe the route the bread takes during the course of digestion." The following is typical of the response given by these eight students:

> "The food must first be broken down with the help of your teeth. Then once in the stomach the acids break it down even more. After the stomach it enters the intestines and there are more chemicals that break down the food that is to be digested."

Although this might seem like a reasonable response given the question students were asked to respond to, the concept map used to guide scoring of students' written explanations displays quite a different interpretation; students were expected to explain how carbohydrates in bread are converted into useable energy and other by-products through cellular respiration (see Figure 12).

Because 8 of the 12 students interpreted the conceptual integration task as "route the bread travels during digestion" as opposed to the intended or expected interpretation—cell respiration—these 8 students scored zero. Consequently, the digestion of bread question is not included in any further analysis.

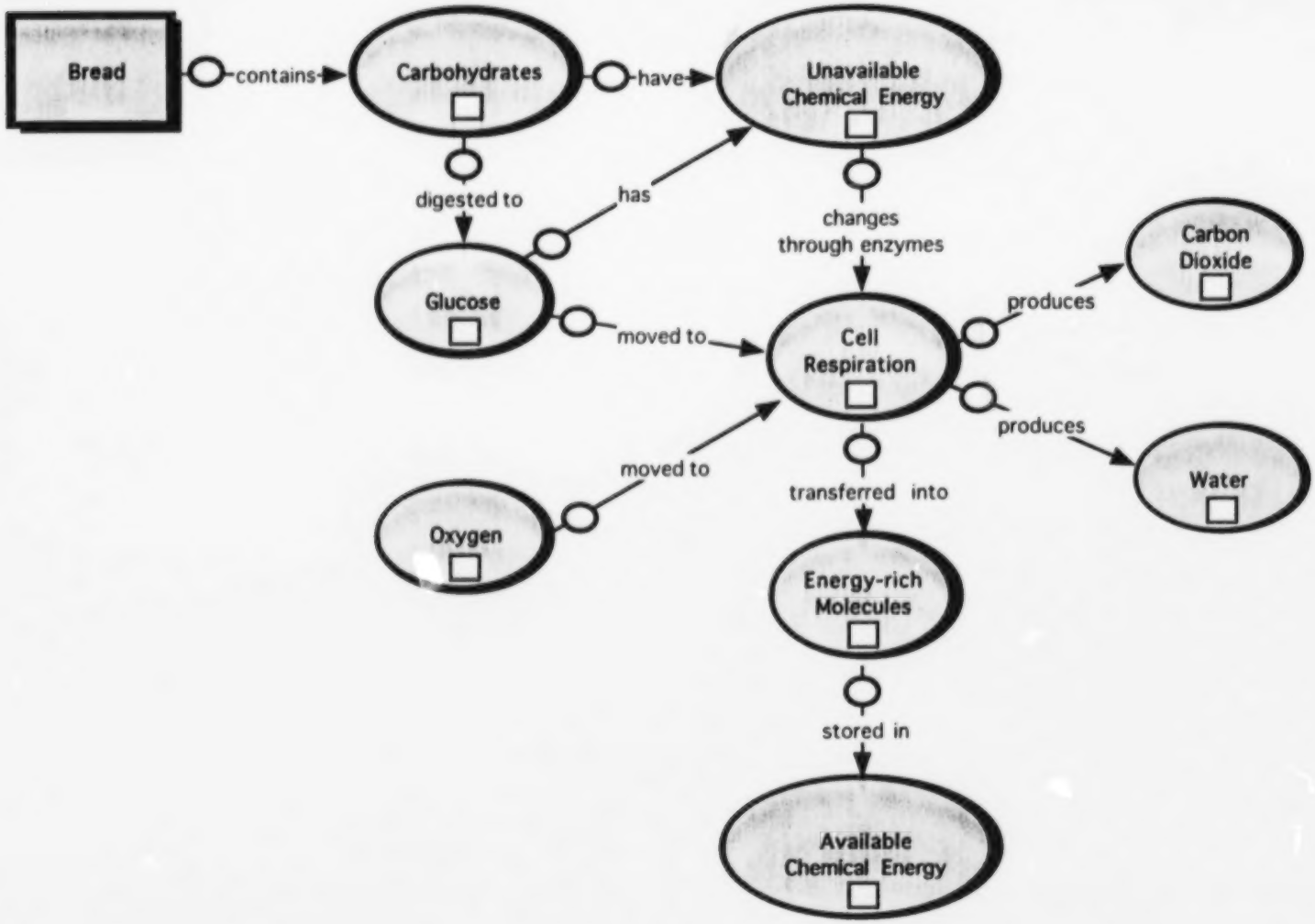


Figure 12. Concept map for scoring the Conceptual Integration task, Digestion of Bread (adapted from Connecticut Common Core of Learning Assessment Project, 1991).

BEST COPY AVAILABLE

Growing Plants. Students were asked to "describe the possible forms of energies and types of materials involved in growing a plant and explain fully how they are related." Unlike the digestion of bread question, students formulated responses that matched, to some extent, the expected interpretation as represented in the expert concept map (see Figure 13). The sole exception was one AP student who explained, "When I first saw the question I thought we would have to discuss how the seed grows up into the root. And then photosynthesis would come later on so that kind of threw me off track." By focusing primarily on the development of the seed, and to a lesser extent on photosynthesis, this student scored quite low.

An examination of students' written responses indicated that, in general, students, regardless of current science enrollment, mentioned water and sunlight (see Table 2). Nine students who mentioned photosynthesis and one student who mentioned chlorophyll were given credit for photosynthetic pigment (e.g., chlorophyll). Further, 50% of the students mentioned carbon dioxide or oxygen; only four students mentioned both. Although the question specifically asked for "forms of energy," geology and human biology students did not mention chemical energy or energy-rich molecules. Nor did these students mention new materials or sugar in their explanation.

Interviews with students suggest that, with the exception of AP biology student 8, the written explanations are reflective of the concepts students associate with the topic "growing plants." (Student 8 misinterpreted the question; see above.) In general, students wrote about the concepts they felt were related to growing plants and opportunities to elaborate resulted in, at most, one or two more concepts being mentioned; most students could not elaborate on their written response (see Table 2). AP biology students mentioned more concepts in their written explanations and added a larger number of concepts in their interviews than the students in the other classes. This finding may, in part, reflect current science enrollment. AP biology students had recently studied photosynthesis; not so human biology and geology students. As one geology student indicated: "The last time we learned about growing plants was in biology, um, I took biology in 9th grade and I'm now a senior and I've already taken like four other science courses . . . we haven't done that since biology so it just dealt with memory, and that's about all I could remember."

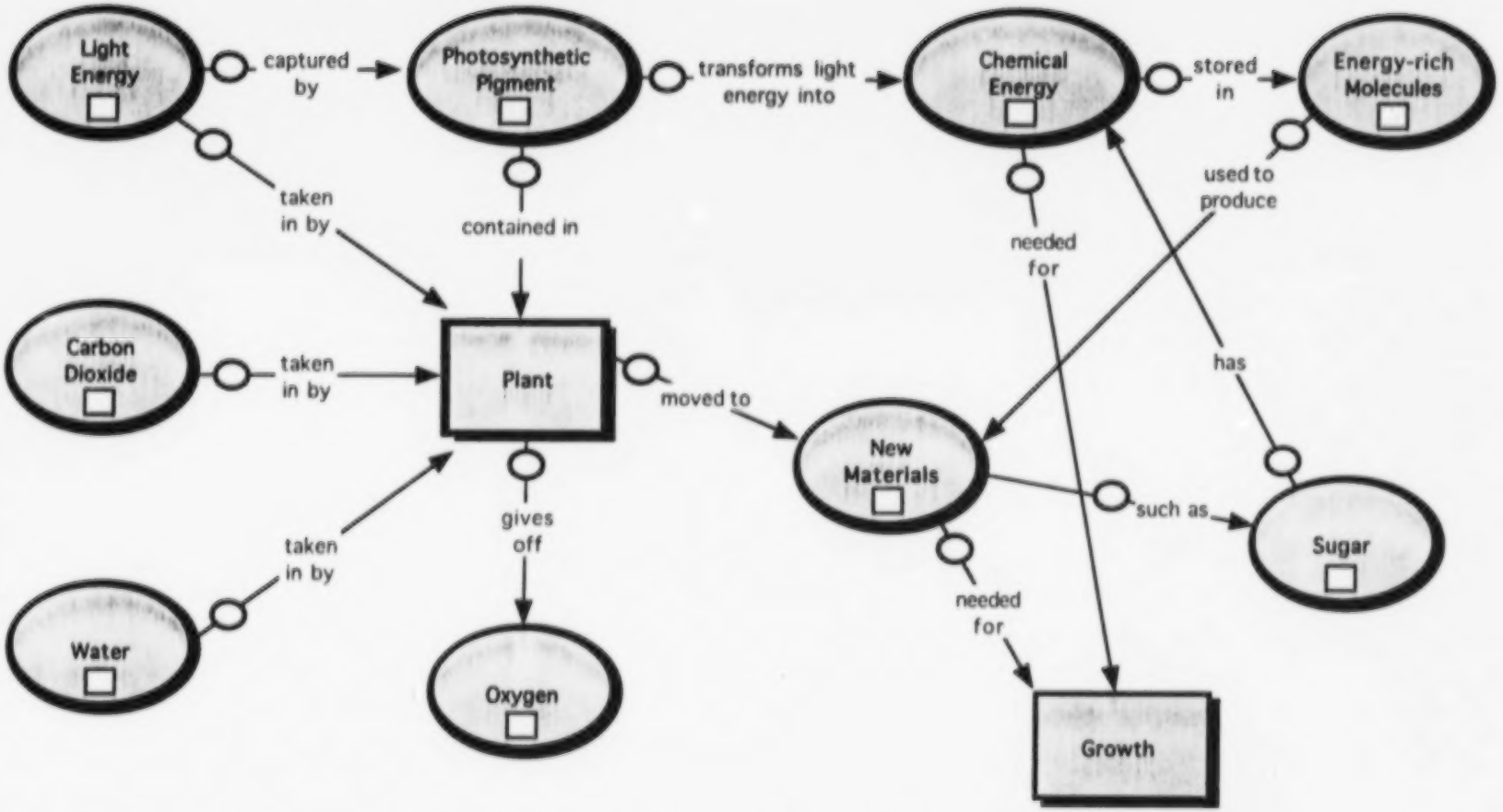


Figure 13. Concept map for scoring the Conceptual Integration task, Growing Plants (adapted from Lomask, Baron, Greig, & Harrison, 1992).

BEST COPY AVAILABLE

Table 2

Concepts Mentioned in Written Explanations (x) and Interviews (o) in Response to the Conceptual Integration Task, Growing Plants

Science class/ID	Light energy	Carbon dioxide	Water	Photo-synthetic pigment	Oxygen	Chemical energy	Energy rich molecules	New materials	Sugar
Human biology									
1	x		x	o					
2	x	x	x	x	x				
3	x		x	x	x				
4	x		x	x	x				
AP biology									
5	x	x	x	x		x	o	x	x
6	x	x	x	x		o	o	x	x
7	x	x	x	x	x		o		o
8	x		x		o	o	o	o	o
Geology									
9	x	x	x	x	x				o
10	x		o	x	o				
11	x		x	x					
12		x		x	x				

Evaluation of student responses was based on the number of concepts mentioned and the explanation of how these concepts are related. A comprehensive, coherent explanation of the "forms of energy and types of materials" involved in plant growth and how they are related revolves around a discussion of inputs, processes, and products. The plant takes in water, light, and carbon dioxide. Through the process of photosynthesis light energy is converted into chemical energy used to produce new materials such as sugar needed for plant growth. In addition oxygen is given off.

In evaluating students' explanations, then, we focused on the expressed relationships among the inputs, processes, and products. Although five qualitatively distinct levels could be described, from nonexplanation (Level 1) to ideal explanation (Level 5), in this sample we found examples for Levels 1 through 4 only (see Figure 14). The best explanations came from 2 AP biology students; one provided a good explanation and the other a partial explanation. Of the remaining 10 students, 5 presented fragmented explanations and 5 presented nonexplanations.

Level 1 (Nonexplanation)

"A growing plant gets energy from the rays of sunlight which enables the plant to make chlorophyll."

Level 2 (Fragmented)

"Growing plants is an everyday occurrence that is often taken for granted. Plants use photosynthesis to help themselves grow. Photosynthesis involves taking in sunlight and absorbing water through the roots of the plant."

Level 3 (Partial)

"Well when one plants a seed, the right conditions have to be met for the plant to grow. It needs the proper amount of moisture and the right nutrients have to be present in the seed. When the seed breaks through the soil it would be exposed to light where the plant gets its energy from through a process of photosynthesis. Light hits the plant where the chloroplasts (the green things) turn it into ATP adenine triphosphates. To get to that stage there are many cycles, the light reaction, the dark reaction, the calvin cycle. Photosynthesis is usually the same in all plants except for cacti where due to the lack of water they only gather or give of CO_2 or O_2 at night. I really don't remember."

Level 4 (Good)

"The first form of energy to consider when examining a growing plant is that of light. This light excites the electrons of the chlorophyll in the plants' chloroplasts. As the electrons fall back to their original state they release energy that can convert [ADP + P] to ATP. This light process of photosynthesis is essential to energy gathering by the plant. Additional forms of energy are gathered by the interactions of other chemicals such as H_2O, CO_2, $NADP^+$, etc. The final product of the photosynthetic process is the production of glucose for the plant. This energy allows the plant to continue its life processes. Additionally, nutrients from the soil (e.g., minerals) are absorbed into the roots. Those nutrients make the root hypertone to the surroundings and allow for the intake of water to aid in the process of photosynthesis and ultimately in the process of energy gathering."

Figure 14. Examples of student explanations for the Conceptual Integration task, Growing Plants.

We compared this qualitative evaluation of their written explanations to performance scores (level of understanding based on size and strength scores) assigned by teachers using the concept map scoring system. We expected those students who provided the most complete, coherent explanations (e.g, good or ideal) to have the highest level of understanding (i.e., an overall score of 4 or 5). Results indicate that the quality of the students' explanations differed slightly from the level of understanding (see Table 3). All students who received an overall score of 1 (size = small and strength = weak or medium) provided nonexplanations. Students who scored 2 (size and strength

Table 3

Comparison of Level of Understanding and Quality of Written Explanation for the Conceptual Integration task, Growing Plants

Science class/ID	Overall score	Level of understanding		Quality of explanation
		Size	Strength	
Human biology				
1	1	Small	Weak	Nonexplanation
2	2	Partial	Medium	Nonexplanation
3	1	Partial	Weak	Nonexplanation
4	3	Partial	Strong	Fragmented
AP biology				
5	4	Substantial	Strong	Good
6	3	Partial	Strong	Fragmented
7	3	Partial	Strong	Partial
8	1	Small	Weak	Nonexplanation
Geology				
9	2	Partial	Medium	Fragmented
10	1	Small	Medium	Nonexplanation
11	2	Small	Strong	Fragmented
12	2	Small	Strong	Fragmented

= partial and medium or small and strong, respectively) provided nonexplanations or fragmented explanations. Students who scored 3 (size = partial and strength = strong) provided fragmented or partial explanations.

The most notable lack of correspondence between the level of understanding and the quality of explanation arises in the evaluation of the cohesiveness of students' knowledge. Evidence from interviews suggests that the strength scores overestimate students' understanding of the relationships among concepts mentioned in their explanation. For example, consider the following written explanation:

> "Growing plants is an everyday occurance that is often taken for granted. Plants use photosynthesis to help themselves grow. Photosynthesis involves taking in sunlight and absorbing water through the roots of the plant."

During the interview when asked about photosynthesis, the student responded: "The process it uses to take in the sunlight, um, to keep the plant alive. It's its way of feeding." With further prompting, the student responded:

"What does it use as food? Um, I don't know, minerals from the soil, it uses sunlight, it uses the water. It stores them and then just as our body does it passes it throughout, it passes through itself."

The student was given a score of 2, size=small and strength=strong. Describing this student's knowledge as "well connected" is clearly an overestimate as evidenced by the response given during the interview.

In summary, then, students write what they know and accurately display their level of understanding in their written explanations. Opportunities to elaborate resulted in little additional information, which would indicate that students knew more than they wrote. For most of the students, their understanding was rudimentary—plants need sunlight and water to grow. Other concepts (e.g., photosynthesis) although mentioned by students in their written explanations could not be explained in any detail. Nevertheless, for one-half of the students, size and strength was described as "partial and strong" or "small and strong." Evidence from the interviews suggests this is an overestimate of what students understand about the "forms of energy and types of materials" involved in growing plants. It appears from the analysis that the concept map scoring system, as presently designed, is not reflective of students' knowledge structures but rather serves as a checklist of terms students associate with a given topic.

Blood Transfusions. Students were asked: "For what would you want your blood checked if you were having a transfusion. Explain fully why blood should be checked for these things." The expected response would contain references to the possibility of an immune reaction (blood compatibility) and acquiring an infection (see Figure 15). In developing a response students were expected to give examples of three blood-transmitted diseases (HIV, syphilis, and hepatitis B). In addition "teachers expected their students to understand the reaction between RBC [red blood cell]-surface antigens and naturally circulating antibodies as the basis for blood compatibility" (Lomask et al., 1992, p. 11).[2]

[2] There are two antigens that may be present on the red blood cells (A and B). An individual may have one of these antigens (i.e., type A or type B), or both (type AB), or neither (type O). Within the plasma of an individual, there are antibodies to the antigens that are not present on that individual's red blood cells. Thus, for example, type A blood has antibody B. Type AB blood has no antibodies because both antigens are on the red blood cells; Type O blood has A and B antibodies. For a recipient to receive blood from a donor, the recipient's plasma must not

An examination of students' written responses indicates that the most commonly mentioned concepts were HIV (92%), blood type (75%), and diseases (58%). Rejection and hepatitis were mentioned by 33% of the students. Two students mentioned sexually transmitted diseases and were given credit on the concept map for mentioning syphilis, although other viral infections such as AIDS, gonorrhea, and herpes are also transmitted this way (Mader, 1990). Concepts that are, for the most part, learned within the context of health or science class (e.g., Rh factor, antigens, antibodies) were mentioned at most by one student (see Table 4).

Table 4

Concepts Students Mentioned in Written Explanation (x) and Interview (o) for Conceptual Integration Task, Blood Transfusions

Science Class/ ID	HIV	Disease factors	Hepatitis B	Syphillis	Type of antigen	Blood type	A,B,O	Rh factor	Rejection	Antibodies
Human biology										
1	o	x				x	o			
2	x	o	x			x	o		o	
3	x	o	x			x	o		o	
4	x									
AP biology										
5	x	x		x		x	o		x	
6	x	x	o			x	o		x	
7	x	o		x		o	o			o
8	x	x				x	o	x	x	o
Geology										
9	x	x	x			x	x		x	
10	x	x				o				
11	x	o	x			x	o		o	
12	x	x				x	o		o	

have an antibody that would cause the donor's cells to agglutinate. Agglutination of red blood cells can cause the blood to stop circulating resulting in death.

Another important antigen in matching blood types is the Rh factor. Persons with this particular antigen on the red blood cells are Rh positive; those without it are Rh negative. Rh negative individuals do not normally make antibodies to the Rh factor, but they will make them when exposed to the Rh factor. This is particularly problematic for Rh negative mother passing antibodies that will attach the red blood cells of an Rh positive fetus (Mader, 1990; Starr & Taggart, 1989).

When given the opportunity during their interview to elaborate or say more about each of the topics, 11 of the 12 students provided explanations that included additional concepts. The majority of students who did not mention diseases, rejection, or A, B, O in their written explanation, mentioned these concepts during their interview (denoted in Table 4 with 0). Students typically responded: "I mean O, A, B, I mean everybody has a different type of blood, and if you have the wrong type of blood it will be rejected by your body system." Only one student responded "A, B, O. I think that's the three," without any further elaboration.

As was the case with the written responses, certain concepts such as antigens and Rh were not mentioned during the interview. When asked specifically: Have you heard of the Rh factor? one student (AP biology) responded affirmatively. "Oh, the rhesus factor. Yeah. That's another thing you'd want that to match. I'm not so sure what that is, like plus or minus, I know that much, based on the rhesus."

Student explanations were evaluated with respect to five levels, as was done with the explanations of the growing plants (see Figure 16). In general, performance was less than optimal; one-half of the 12 students provided nonexplanations and only one student provided a good explanation. AP biology students provided the best explanations; human biology students the poorest.

An examination of scores assigned by teachers using the concept map indicated that students had very little conceptual knowledge as evidenced by the size dimension (see Table 5). Of the 12 students, 6 mentioned 12% to 35% of core concepts (size=small) and 3 mentioned 35% to 58% of core concepts (size=partial). Nevertheless, based on the concept map scoring system, students were judged to have very well integrated knowledge of those few concepts they did know (10 of 12 students scored strong on strength). This finding contrasts with the quality of students' explanations (see Figure 16); one-half of the students provided nonexplanations.

A closer examination of student written responses suggests that the scoring system as presently designed may overestimate levels of understanding in two respects. First, one-half of the core concepts are learned in contexts outside science class (HIV, disease, blood type, hepatitis B, and syphilis). Second, the relations among the concepts are at the level of

Level 1 (Nonexplanation)

"Well first of all I would want it checked to see if the blood was the same blood type as mine. Then I would want it checked for diseases."

Level 2 (Fragmented)

"I would want blood to be checked for Aids and all other illnesses and assorted genetic problems such as sickle cell Anemia. Also Blood type should be checked to ensure a good match. Sickle cell blood cells can't cary O_2 as easily as regular cells and can become lodged in capillarys. Aids can be spread through blood and has killed many people. Most other blood related illnesses can also be transfered which is why they should be checked."

Level 3 (Partial)

"The most serious contaminator of blood would be AIDS. I would want my blood transfusion to be checked for HIV so that I could be sure that I would not be contracting the virus through the blood.
Additionally, I would want the blood transfusion to be checked so that I could know that it is compatible with my blood type.
Finally, I would want the transfusion to be checked for all types of other diseases that could be tranmitted through blood. These include sexually transmitted diseases as well as other diseases that are transmittable through bodily fluids.
It is important that blood be checked for these things because certain diseases can be easily transmitted through blood. It is important to check the blood type to be sure that it is compatible with the rest of my blood."

Level 4 (Good)

"Blood is the primary means of transportation within the human body, and therefore comes in contact with almost all of the body's vital organs. It is therefore imperative that the blood one receives during a transfusion is not harmful in any way. There are several characteristics which make blood harmful to the well-being of the body. Blood type is one of the most commonly known of these characteristics. The blood one receives must accord with the blood type one carries because otherwise the blood may be treated as a foreign substance. Blood must also be tested for whether it is Rh^+ or Rh^-, for the acquisition of Rh^+ blood by an Rh^- person can result in serious illness.
The diseases which have abounded over the recent years have made other checks necessary. Blood must be checked for its HIV content in order to ensure that AIDS will not be spread.
The white blood cell level is also a critical feature of blood which has been disturbed in some people, by the atomic bomb incidents. If one receives blood with a low white blood cell level, one's vulnerability to diseases is greatly increased.
It is thus imperative that utmost care be taken to check blood before a transfusion."

Figure 16. Examples of student explanations for the Conceptual Integration task, Blood Transfusions.

examples and not processes or underlying causal mechanisms. Notice in the Blood Transfusions concept map (see Figure 15) that three connections use the term "is checked for" and six use the term "such as." In the context of this question, students can appear to know or understand a considerable amount

Table 5

Comparison of Level of Understanding and Quality of Written Explanation for the Conceptual Integration Task, Blood Transfusions

Science Class/ID	Overall score	Level of understanding		Quality of explanation
		Size	Strength	
Human biology				
1	2	Small	Strong	Nonexplanation
2	1	Small	Medium	Nonexplanation
3	1	Small	Medium	Nonexplanation
4	1	Irrelevant	Strong	Nonexplanation
AP biology				
5	3	Partial	Strong	Partial
6	3	Partial	Strong	Fragmented
7	2	Small	Strong	Fragmented
8	3	Partial	Strong	Good
Geology				
9	4	Substantial	Strong	Partial
10	2	Small	Strong	Fragmented
11	2	Small	Strong	Nonexplanation
12	2	Small	Strong	Nonexplanation

without ever expressing the interdependent nature of antigens and antibodies and how this is related to the rejection of incompatible blood types.

In summary, students' written responses represented for the most part what students know about necessary checks prior to a blood transfusion. Students know that blood should be checked for diseases, although they sometimes referred to genetic diseases as blood transmitted. They are also aware that there are different blood types, but they do not have a clear understanding of what blood type is, how it is checked, or why it is checked. During the interviews, students mentioned one or two more concepts but could not provide a coherent explanation of those concepts. Characterizing student understanding as "small and strong" or "partial and strong" is only partially accurate. Size (number of concepts) is very small, but strength (connections among concepts) is not strong. Rather, students appear to have a shallow understanding of the types of things blood should be checked for prior to a transfusion and the necessity for checking for each of these.

Discussion

The Conceptual Integration task is intended to measure what students have learned after high school science regardless of the particular courses they have studied. These open-ended tasks requiring students to decide which concepts are important and how they are related provide an opportunity for students to display their level of understanding. Nevertheless, the open-ended nature of the task sometimes left students unaware of the intended interpretation of the question, and the scoring system, as currently designed, does not allow for alternative but reasonable interpretations. This was particularly problematic for the digestion of bread question.

For the growing plants and blood transfusion tasks, interviews revealed that students' written responses were accurate representations of their knowledge and understanding of each of the topics. In general, regardless of the question, students were successful at listing ideas they associated with the question and less successful at providing in-depth explanations. Students' written responses were evaluated on the basis of a match with the "expert" concept map. The scores derived from these maps did not match the evaluation of students' understanding of the topic (from the interviews) or the quality of their explanations. The most apparent discrepancy was with the strength component (representing the number of valid connections among core concepts) of the scoring that tended to overestimate the coherence of the students' knowledge. Some students who provided little more than a list of ideas were scored as having strong, interconnected knowledge.

Student performance was, on average, higher on the blood transfusion task than on the growing plant question. This may be a reflection of: (a) the length of time between studying the concepts on the assessment and current science enrollment, (b) the nature and depth of understanding required to provide the "expected" response, or (c) some combination of these. Proficient performance on the growing plants question required students to reflect on the relation between inputs, outputs, and the process that links these. Students who focused primarily on inputs—plants need water and sunlight—received low scores.

In contrast, successful performance on the blood transfusion question was much more dependent on learning opportunities outside the classroom

(e.g., HIV) and less on content knowledge learned in science class (e.g., relation between antigens and antibodies). Further, the strength score was unduly weighted on the ability to cite familiar examples of a more general concept rather than on an understanding of causal relations. Students were given credit for disease types and three examples (HIV, syphilis, hepatitis B), which comprised 4 of the 10 possible concepts. The links between these concepts were "such as." A statement "I would want blood checked for diseases such as HIV, syphilis and hepatitis B" would be scored 4/10 size and 3/3 strength. Level of understanding would be partial and strong. This judgement is clearly an overestimate of what this student knows about blood checks prior to a transfusion.

From our analysis, we conclude that the concept map format can be a good way to represent students' knowledge in a domain. By providing a visual display of the relevant concepts and their interrelations, students have an example of quality performance by which to judge their own understanding. However, unless proficient performance displayed by the concept map requires inferences or reasoning about subject matter relations or causal mechanisms reflective of cohesive, in-depth knowledge, then it serves as little more than a checklist of words and in the end misrepresents (overestimates) students' level of understanding.

Component Identification

Tasks of this type, of which Electric Mysteries is an example, require students to reason with their conceptual knowledge to identify an unknown entity. For the Electric Mysteries assessment situation students engage in a cyclical process of hypothesis testing and refining to identify the circuit components enclosed in each of six boxes. Using their knowledge of what constitutes a circuit, and the impact of changing various components in a circuit (e.g., adding a second bulb), students test out their hypothesis by observing a bulb connected in a circuit external to the box. For example, if the bulb is dim when connected in a circuit to one of the boxes, students might reason that there is a battery and a bulb in the box. If the bulb is very bright, students might reason that there are two batteries in the box.

Task

This task has two parts (see Figure 17). Part I serves to orient students to the equipment (batteries, bulbs, wires with clip leads, and mystery boxes) and the task. To this end, students are asked to create a circuit and draw a picture of it. Then, students are asked to determine whether a battery or a wire is in a box with a question mark on it. They are prompted to connect the question mark box in a circuit with a bulb to determine what is inside.

Part II of the task asks students to determine the contents of each of six "mystery" boxes A through F from a list of five possible alternatives (see Figure 18). Students are presented with two batteries, two bulbs and five wires to connect circuits to the six boxes. Two of the boxes have the same thing (Box B and F each have a wire). All of the others have something different (battery and bulb, two batteries, nothing, or a bulb).

Scoring

Student performance is scored on the basis of (a) identification of the contents of each box, and (b) the circuit used to arrive at the answer (see Figure 19). If the student correctly identifies the contents of a box and draws a circuit that legitimately leads to that conclusion, he or she is credited with one point. If the student fails to correctly identify the contents and/or fails to demonstrate the use of an adequate circuit, no credit is assigned. The maximum possible score is 6 (one for each box). The initial part of the task (build a circuit and determine what is in the question mark box) is not scored as these questions serve to orient the student to the task.

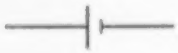
Data Collection

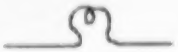
Thirty-one students from three classrooms (two in one school and one in another) in an urban California school district were interviewed. All students had recently (within a 2-week period immediately preceding conduct of this investigation) completed the same 8-week unit of study on electric circuits. Prior to conducting the investigation, instructions were read aloud to each student and the equipment was introduced. Students were asked to talk aloud as they worked on the task. Prompts of "What are you doing, why are you doing it, what are you going to do?" were continuously provided as students

You have some batteries, bulbs, and wires in front of you for doing some experiments. All the wires are the same. They are just different colors. They have clips on the end so you can connect things.

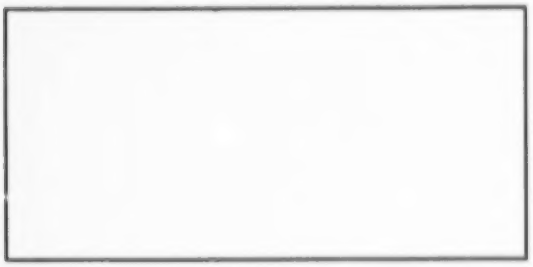
1. Connect one battery, one bulb, and wires so the bulb lights.

Draw a picture in the box that shows what you did. If you like use symbols like these:

Battery symbol: 

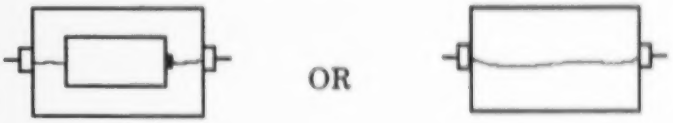
Bulb symbol: 

Wire:

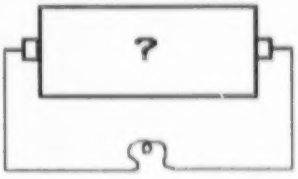


2. Figure out what is in the mystery box labelled with a question mark (?).

The box has either a battery or a wire inside:



To help figure out which one is in it connect it in a circuit with a bulb:



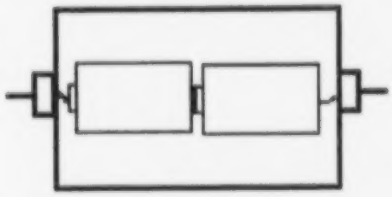
Fill in the answer:

The "?" box has a _____ in it.

Figure 17. **Component Identification task, Electric Mysteries Part I (adapted from Shavelson, Baxter, & Pine, 1991).**

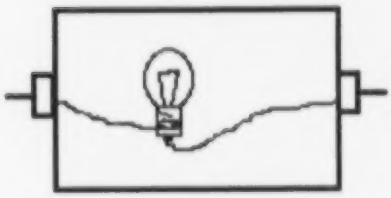
Find out what is in the six mystery boxes A, B, C, D, E, and F. They have five different things inside, shown below. Two of the boxes have the same thing. All of the others have something different inside.

Two batteries:

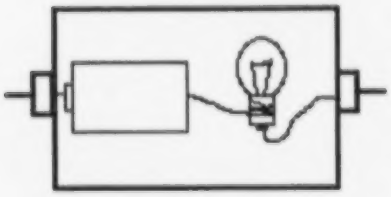


A wire:

A bulb:



A battery and a bulb:

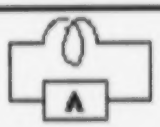
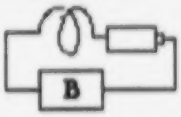
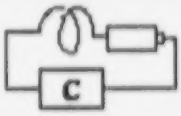
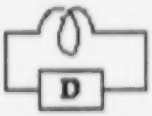
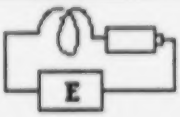
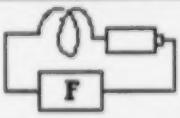


Nothing at all:

For each box, connect it in a circuit to help you figure out what is inside. You can use your bulbs, batteries, and wires any way you like.

When you find out what is in a box, fill in the spaces on the following pages.

Figure 18. Component Identification task, Electric Mysteries Part II (adapted from Shavelson, Baxter, & Pine, 1991).

MYSTERY BOX	ANSWER	CIRCUIT	SCORE
A	Battery and Bulb	Bulb only 	
B	Wire	Battery and Bulb 	
C	Nothing at all	Battery and Bulb 	
D	Two Batteries	Bulb only 	
E	One Bulb	Battery and Bulb 	
F	Wire	Battery and Bulb 	

TOTAL SCORE

Figure 19. Scoring system for the Component Identification task, Electric Mysteries (adapted from Shavelson, Baxter, & Pine, 1992).

conducted the task. Interviewers recorded the sequence of circuits students used to test each box as a measure of their problem-solving strategy.

In addition, two questions were inserted at key points in the investigation to elicit information about students' task-specific conceptual understanding. After students built a circuit in Part I of the task, they were asked: "Can you tell me what a circuit is? Can you tell me how a circuit works?" One question on polarity was used, in addition to the question about circuits, as evidence of students' conceptual understanding. After completion of Part II of the task, students were shown a battery, bulb, and two wires connected in a circuit. They were asked to predict what would happen if a second battery was added in series to this circuit. If opposite poles (positive and negative) are connected, the bulb will light brighter than a bulb in a circuit with one battery; if like poles (positive with positive or negative with negative) are connected the bulb will not light. The interviewer then added the battery to the circuit with like poles connected and when the light bulb failed to light, the students were asked to explain this outcome.

Before beginning Part II of the task—determine what is in each of the six mystery boxes—students were asked: "How are you going to go about solving this problem?" This question was used as evidence of students' ability to generate a knowledge-based plan prior to solving the component identification task.

Results

Our analysis focused on the relationship between planning, conceptual knowledge, problem-solving strategy, and performance score. Is there a difference between students who score high and students who score low in terms of: (a) their ability to generate a plan based on subject matter knowledge, (b) the quality of their explanations of task-related concepts (circuit, polarity), and (c) the efficiency with which they solve the problem? Because the component identification task requires students to reason with their knowledge of circuits, we expected to find relations between quality of knowledge-based plan, level of conceptual understanding, efficiency of problem solving and score such that students rank consistently (high or low) on all four aspects. More proficient students (those who score high) were expected to

generate a knowledge- based plan, display greater conceptual understanding, and be proficient in their solving of the task.

Plan. In general, when presented with a task, proficient students seek to understand the nature of the task by generating a representation to test alternative solution strategies. This representation is used to anticipate alternative outcomes to various actions (e.g., testing with a bulb only in the circuit) and to generate next steps based on those outcomes (e.g., testing with a bulb and battery in the circuit). In other words, proficient students think through the problem solution prior to actually engaging in manipulating the equipment. This thinking through, or trial run, is expressed in their ability to generate a knowledge-based plan. Students without a strong conceptual understanding are unable to generate an adequate representation of the task that would allow them to reason through the problem in the abstract. Rather, they rely on concrete feedback from manipulating task-related equipment to suggest next steps or alternative strategies.

Results of our analysis indicate that all students who scored 3 or higher offered a plan. The plans were task specific and focused almost solely on the impact of connecting the bulb in a circuit to the box. Students typically stated they would use "bulbs to see if it lights" or "bulb will say if battery in the box." None of these high scoring students explicitly stated they would first test each box with a bulb to determine if there was a battery in the box, and then they would test each box with a battery and bulb to determine if there was a wire, a bulb, or nothing in the box. Students with a score of 2 were much vaguer often saying "just like we did." This statement is in reference to procedures they were prompted to use in Part I of the task; students were cued to connect the question mark box in a circuit with a bulb to determine the contents. Other students who scored 2 focused on the equipment they were going to use "batteries, bulbs, and wires" and not on how this equipment would help them determine the contents of each box. Students who scored 1 or zero did not offer a plan but rather began solving the problem saying "I am going to hook it . . ." as they connected a wire to the box.

In summary then, proficient students (those with the highest scores) generated a reasonable plan. Those students with the lowest scores began attacking the problem at hand, hooking up various pieces of equipment trying to determine the contents of each box. Their lack of circuit knowledge

precluded their thinking through their solution strategy prior to beginning to solve the problem, and consequently they did not offer a knowledge-based plan for approaching the task.

Explanations. Conceptual knowledge was judged on the basis of student explanations of (a) how a circuit works, and (b) why a bulb fails to light when a second battery is added, in series, to a circuit. It was expected that students who scored high on the task would show evidence of conceptual understanding through the quality of their explanations.

In response to the question "Can you tell me what a circuit is?" students who scored 5 or 6 typically described a closed system consisting of wires, battery (for power or energy), and a bulb. Further, they explained that the circuit works by "a flow of electricity that goes in a complete circuit 'cause it has to travel in a complete circuit from a battery to a bulb and back, so like this, that's a pathway." Students who scored 3 or 4 also described a closed system and mentioned "pathway" in their explanation of a circuit. However, when prompted to "tell me how a circuit works" students would respond, "Energy comes from each side of the battery to wires to inside the bulb," betraying an alternative conception of energy flow in a circuit. Students who scored 2 were much more varied in their responses. Some gave responses similar to students who scored 3 or 4, stating, "Energy from the negative side flows to positive side and they interact." Others gave less adequate explanations such as "battery pass on electric to the light bulb" or "energy, electricity flows like a pathway." For students who scored 1 or zero, two of the five responded "I don't know." The other three students mentioned the battery being a source of electricity or energy. Although students generally offered nonelaborated explanations, the quality of explanations did vary by score level; those who scored highest provided the best explanations, relatively speaking.

Student explanations of why a bulb failed to light when a second bulb was added in series also varied by performance score. The majority of students who scored 5 or 6 mentioned the need to have positive with negative otherwise "it wouldn't flow through." Six of the 12 students who scored 2, and 3 of the 5 students who scored 1 or zero couldn't provide any explanation. Four students who scored 2 mentioned the "ends are the wrong way." All other students who scored 2 or less gave inadequate responses such as "battery loses energy, it splashes out."

In summary, student explanations, while not elaborative, varied by performance score. Those who scored high (5 or 6) provided better explanations than those who scored low (0 or 1). Moreover, the quality of students' explanations was consistent for the circuit and polarity questions. Students who provided good explanations of what a circuit is also provided good explanations of why the bulb failed to light when a second battery was added, in series, to the circuit.

Strategy. Principled problem solving is reflective of knowledge organization and structure. If students know/understand the impact of changing various components in a circuit, then they will only add those components that are most informative and, further, will do so in a reasoned, principled fashion. It was expected that students who scored high would demonstrate the use of more efficient strategies and students who scored low would demonstrate use of relatively less efficient strategies. To examine this aspect of student performance, the sequence in which students built circuits and connected them to boxes was categorized in terms of efficiency of solution strategy. The least efficient strategy is a trial-and-error approach. The most efficient strategy is—connect each box with a bulb to determine which of the six boxes has a battery. Then connect the remaining boxes with a battery and bulb.

Students who scored 5 or 6 displayed a principled approach to solving the problem. Typically they connected each box in a circuit with a bulb. If this circuit did not provide confirmatory evidence as to the contents of the box, the box was then connected in a circuit with a battery and a bulb. Few extra circuits were tried except when the student checked to see if the bulb in the circuit did not light because the batteries were connected with like (e.g., positive to positive) ends rather than opposing (positive to negative) ends. Students who scored 3 or 4 demonstrated more trial-and-error in their approach, tended to repeat circuits or try many different circuits, tested some boxes with a battery only in the circuit, and failed to check for the direction the batteries were facing when they did use a battery and bulb in the circuit to test the box. Only 4 of the 12 students who scored 2 used a battery and bulb in a circuit to test the boxes. The remaining 8 students and all of the students who scored 1 or zero tested all boxes with a bulb only or a battery only. Students who scored 1 typically tried very few circuits.

Results of this analysis indicate a relationship between student performance score and problem-solving strategy. Those students who scored high drew on their knowledge of circuits to determine the contents of each box. In doing so, they tried specific, informationally-rich circuits in a systematic fashion—first bulb, then battery and bulb. Those students who scored low reflected their lack of circuit knowledge in their performance; they tried very few meaningful circuits and never tried a circuit with both a battery and a bulb in it. Indeed some of these students tried circuits with only a battery in them in an effort to determine if there was a bulb in the box. This strategy would only work if light could be seen through the box, which it could not, in this case. For other students, fragmented or partial understanding of circuits was reflected in their trial- and-error approach.

Summary

The Component Identification task examined in this study, Electric Mysteries, requires students to reason with knowledge of circuits to identify the circuit components enclosed in each of six boxes. Scoring was based on identification of the contents of each box and the legitimacy of the circuit used to do so.

In general, students who scored high (5 or 6) on the assessment, were able to (a) describe a plan for carrying out the investigation, (b) express through their explanations an understanding of the conceptual knowledge of circuits, and (c) demonstrate an efficient, principled approach to solving the problem. Those students who scored low (0 or 1) were not able to describe a plan, offer an adequate explanation for task-related concepts, nor could they approach the problem with any knowledge-based reasoning. For those students with partial or developing knowledge of circuits (scores of 2, 3, or 4), some students showed strength in some areas (e.g., principled problem solving or efficiency) but poor understanding in others (e.g., ability to generate a knowledge-based plan). Vice versa for other students. Unlike students who scored at the extremes of the score scale, performance for these students was not consistent across the various dimensions of proficient performance examined in the context of this assessment task.

In conclusion, the Component Identification task allows students to demonstrate their level of knowledge and understanding. Moreover, student

scores are linked to dimensions of proficient performance (plan, conceptual understanding, principled strategy). Further, the scoring system provides feedback to teachers and students regarding the procedural steps required to conduct the task. For example, by looking at the score form, it can be seen that to determine there is a bulb in a particular box, students would need to connect it in a circuit with a battery and bulb. Students are not, however, provided information as to the characteristics of the external bulb (dim, bright) which can be used to indicate what is inside the box—the bulb is dim, therefore there must be another bulb in the box. The feedback provided by the scoring system (answer and circuit) may not be informative without the accompanying rationale or explanation for using a particular circuit.

IMPLICATIONS FOR THE DESIGN OF ALTERNATIVE ASSESSMENTS

Efforts to develop alternative measures of student achievement in science commensurate with the educational reform agenda focus on creating assessments that display how students use their knowledge to reason and solve problems in contextually relevant situations. These developmental efforts are based primarily on informed intuitions and assume that tasks can be created, administered, and scored to obtain reliable and valid information about students' higher order thinking skills. The purpose of this study was to test this set of assumptions. Specifically, this study analyzed the knowledge and cognitive processes that students exhibit on science tasks that are already part of innovative assessment programs—a pilot formative program in Connecticut and a more defined trial in California. Our intent was to characterize the kinds of performances actually elicited from students and describe how this performance differs among students at various levels of proficiency.

Three types of assessment tasks were selected—Exploratory Investigation, Conceptual Integration, and Component Identification—each varying with respect to grade level, prior knowledge, stage of development, and purpose. Exploratory Investigation tasks were designed to assess 12th-grade students' ability to reason with models to describe the flight of a maple seed. Conceptual Integration tasks were designed to assess 11th-grade students' understanding of the interrelationships among concepts in biology, physics, and chemistry. Component Identification tasks were designed to assess 5th-

grade students' ability to apply their knowledge of circuits in a novel problem-solving situation.

Scoring systems varied by type of assessment task. For the Exploratory Investigation task students' performances were scored on the basis of a match between scoring criteria (e.g., number of observations) and written response. Each component of the task (observation, explanation, reflection, and application) was scored independently. Concept maps were used to guide scoring of the Conceptual Integration tasks. Students' written explanations were examined for their correspondence with the "expert" concept map. The intent was to visually represent students' knowledge structures (concepts and their interrelations) for each topic. Students' performances on the Component Identification task were evaluated on the basis of two aspects of the students' work: the procedure used (e.g., tested the box in a circuit with a bulb) and the resulting identification of the unknown entity (e.g., two batteries).

The verbal protocols and other analyses of students' performances in each of these assessment situations were guided by general dimensions of problem solving on which more or less proficient students differ. Protocol analysis in conjunction with observations of student performance, evaluation of student test booklets, and an examination of the scoring criteria provide an empirical basis for linking performance scores with level and kind of reasoning and understanding. In taking this approach, we anticipated that students who score high on the assessment should exhibit some characteristics of proficient performance (ability to plan, reason, explain, draw inferences, systematically solve problems, and monitor their own performance) if the task requires students to engage in higher order thinking skills.

From these analyses we have begun to formulate critical features of assessments and scoring systems necessary to ensure that appropriate higher order thinking skills are being tapped. These features have implications for task specification and for scoring method specification. Certain task and scoring system features implied by this study are described below.

- Tasks should be procedurally open-ended, affording students an opportunity to display their understanding. The tasks examined in this study refrained from providing step-by-step instructions for students to follow. Rather, students were presented with a problem and prompted to pursue their own line of inquiry. This inquiry may be knowledge-based or driven primarily by surface features of the task.

For example, students with knowledge of electric circuits approached the Component Identification task systematically, connecting each box in a circuit with a bulb to determine if there was a battery inside the box. The trial-and-error approaches of other students signal their lack of knowledge of circuits. Differential approaches to solving the problem are masked when procedural instructions are provided.

- Tasks should draw on subject matter knowledge. Prior knowledge of generally familiar facts (e.g., plants need water to grow, plants need sunlight to grow) is not sufficient for generating complex arguments to support an explanation of naturally occurring phenomena. Rather, sophisticated reasoning and thinking are dependent on an understanding of the causal mechanisms that relate facts or concepts (e.g., energy from sunlight is converted into stored chemical energy during photosynthesis). This kind of understanding develops primarily through focused study of a given phenomenon and, to a lesser extent, through casual exposure more typical of learning opportunities outside the classroom.

- Tasks should provide for the application of knowledge. Recalling specific facts or generating a list of attributes can be accomplished with superficial understanding of a given topic. In contrast, reasoning and problem solving are dependent on knowing when, and under what conditions, specific content knowledge is useful. For example, critiquing a report of a maple seed experiment requires (a) content knowledge (i.e., laws of motion, maple seeds, and experimental procedures), and (b) awareness of the applicability of this knowledge. The degree to which knowledge is connected to its conditions for use determines the adequacy with which this task is completed.

- Scoring criteria should match task expectations. Cognitively rich tasks that ask students to reason with subject matter knowledge demand scoring systems that detail the quality of that reasoning process in a given context. For example, if students are asked to reflect on the value of model-based experiments, then scoring should address the expression of students' reasoning. Checking for the presence of a generic list of advantages and disadvantages of models does not necessarily provide information about students' understanding of, and facility with, model-based reasoning. Scoring systems that emphasize easily quantifiable aspects of performance sabotage efforts to measure higher order thinking skills.

- Scoring should be sensitive to the meaningful use of knowledge. If students' performances are not to be over- or underestimated, a link between performance score and level of reasoning and understanding is critical. For example, the completeness and coherence of students' explanations should dominate scoring as opposed to the number of statements or key words included in the response. Focusing on key words when students are asked to explain fails to distinguish those

students with cohesive knowledge from those with fragmented knowledge, a key distinguishing feature of proficient performance.

- Scores should capture the process students engage in. Knowing when and under what conditions knowledge is useful is reflected in problem solving strategies and processes. For example, when given an observational task, proficient students engage in an iterative process of hypothesis generating and testing until pertinent aspects are reported with a high level of confidence. Less proficient students, on the other hand, may list observations with little understanding of disciplined inquiry prerequisite to reliable observation. Performance scores that attend to the product (e.g., number of aspects observed) without explicit consideration of the process (e.g., hypothesize, test, conclude) encourage unwarranted assumptions about students' understanding.

Characteristics or features of assessments such as those just described provide the basis for an assessment development framework that acknowledges proficiency as the ability to use subject matter knowledge to reason and solve problems. The need for empirical justification for inferential bridges from student score to the nature and extent of student learning is apparent in our analysis of a diverse range of alternative assessment practices. Sensitizing test developers to ways in which students' reasoning can be elicited and scored is an important first step toward progress in the development of assessments commensurate with educational goals.

REFERENCES

Baron, J. B. (1991, April). *Beyond the promise: Design, data and measurement in new forms of assessment.* Symposium conducted at the annual meeting of the National Council on Measurement in Education and American Educational Research Association, Chicago,.

Baron, J. B., Carlyon, E., Greig, J., & Lomask, M. (1992, March). *What do our students know? Assessing students' ability to think and act like scientists through performance assessment.* Paper presented at the annual meeting of the National Science Teachers Association, Boston.

Baron, J. B., Forgione, P. D., Rindone, D. A., Kruglanski, H., & Davey, B. (1989, March). *Toward a new generation of student outcome measures: Connecticut's Common Core of Learning assessment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Chi, M. T. H., Bassock, M., Lewis, M.W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145-182.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-305.

Ericsson, A. K., & Simon, H. A. (1984). *Protocol analysis. Verbal reports as data.* Cambridge, MA: MIT Press.

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models.* Hillsdale, NJ: Erlbaum.

Glaser, R. (1992). Expert knowledge and processes of thinking. In D. F. Halpern (Ed.), *Enhancing thinking skills in the sciences and mathematics* (pp. 63-75). Hillsdale, NJ: Erlbaum.

Glaser, R., Lesgold, A. M., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. R. Ronning, J. Glover, J. C. Conoley, & J. C. Witt (Eds.), *The influence of cognitive psychology on testing and measurement* (pp. 41-85). Hillsdale, NJ: Erlbaum.

Green, D. (1980). The terminal velocity and dispersal of spinning samaras. *American Journal of Botany*, *67*(8), 1218-1224.

Linn, R. L., Baker, E., Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*, 15-21.

Lomask, M., Baron, J., Greig, J., & Harrison, C. (1992, March). *ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science.* A symposium presented at the annual meeting of the National Association of Research in Science Teaching, Cambridge.

Mader, S. S. (1990). *Human biology* (2nd ed.). Dubuque, IA: W. C. Brown.

Magone, M., Cai, J., Silver, E. A., & Wang, N. (1992, April). *Validity evidence for cognitive complexity of performance assessments: An analysis of selected QUASAR tasks.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Millar, R., & Driver, R. (1987). Beyond processes. *Studies in Science Education*, *14*,33-62.

Mislevy, R. J. (1989). *Foundations of a new test theory.* Princeton, NJ: Educational Testing Service.

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn.* New York: Cambridge University Press.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *The Journal of the Learning Sciences*, *1*(2), 201-238.

Seter, D., & Rosen, A. (1992). A study of the vertical autorotation of a single-winged samara. *Biological Reviews*, *67*, 175-197.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(4), 215-232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*(4), 347-362.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, *21*(4), 22-27.

Shute, V., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 179-326). New York: Freeman.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Starr, C., & Taggart, R. (1989). *Biology. The unity and diversity of life* (5th ed.). Belmont, CA: Wadsworth.

Ward-Smith, A. (1984). *Biophysical aerodynamics and the natural environment.* New York: John Wiley and Sons.